VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF TECHNOLOGY

(An Autonomous Institute Affiliated to University of Mumbai)

Department of Computer Engineering

Project Report on

ReadWiz: AI Assistant for Mastering any Content

Submitted in partial fulfillment of the requirements of the degree

BACHELOR OF ENGINEERING IN COMPUTER ENGINEERING

By

Nikhil Dhanwani - D12B (12)
Chirag Santwani - D12B (50)
Manraj Singh Virdi - D12B (66)

Project Mentor

Prof. Manisha Mathur

University of Mumbai
(AY 2023 - 24)

VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF TECHNOLOGY

(An Autonomous Institute Affiliated to University of Mumbai)

Department of Computer Engineering

Certificate

This is to certify that the Mini Project entitled **"ReadWiz: AI Assistant for Mastering any Content"** is a bonafide work of **Nikhil Dhanwani (D12B - 12), Chirag Santwani (D12B - 50) & Manraj SinghVirdi (D12B - 66)** submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of **"Bachelor of Engineering"** in **"Computer Engineering".**

Prof. Manisha Mathur

Mentor

Dr. (Mrs.) Nupur Giri

Head of Department

Dr. (Mrs.) J. M. Nair

Principal

Mini Project Approval

This Mini Project entitled **"ReadWiz: AI Assistant for Mastering any Content"** by **Manraj SinghVirdi (D12B - 66), Nikhil Dhanwani (D12B - 12) & Chirag Santwani (D12B - 50)** is approved for the degree of **Bachelor of Engineering** in **Computer Engineering.**

Examiners

1. ………………………………
(Internal Examiner name & sign)

2. ………………………………
(External Examiner name & sign)

Date: 01st April 2024

Place: Chembur, Mumbai

Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

(Nikhil Dhanwani - D12B / 12)

(Chirag Santwani - D12B / 50)

(Manraj Singh Virdi - D12B / 66)

Date: 01st April 2024

ACKNOWLEDGEMENT

We extend our sincere thanks to everyone who contributed to the success of our project, offering our gratitude for their invaluable assistance and guidance.

Special thanks to Dr. J. M. Nair, principal of the V. E. S. Institute of Technology, for her encouragement and support. We are profoundly grateful to Mrs. Manisha Mathur, our esteemed mentor, whose insightful guidance and support were instrumental in navigating through the project. Her expertise has been a cornerstone of our learning and exploration.

Our gratitude also goes to Dr. Nupur Giri, head of the computer engineering department, for her crucial commitment and support, and to Mrs. Priya R. L., the project coordinator, whose advice and inspiration were pivotal to our project's execution.

We thank the teaching and non-teaching staff for their continuous assistance and motivation, which played a significant role in the timely completion of our tasks. Appreciation is also due to our classmates for their support and for sharing valuable information that was essential in completing our project.

Furthermore, we are thankful to all those involved in gathering and providing information, whose contributions have enriched our project. Their help has been indispensable in achieving our goals.

In conclusion, heartfelt thanks are extended to everyone involved in this journey, whose support and guidance lead to the successful completion of this project.

Index

Abstract

This project introduces an innovative approach to text summarization and multimedia content understanding, employing advanced natural language processing techniques and deep learning models. We leverage the Lamini-Flan-T5 model, now named RefinedWiz after fine-tuning on the SAMSUM dataset, to generate concise summaries from input text. Additionally, we integrate a paraphrasing task, enabling users to rephrase text while preserving its original meaning. Our system incorporates chat functionality with PDF documents through the Gemini model, facilitating interactive querying and information retrieval. Furthermore, image and video summary capabilities are included for compact representations of multimedia content. Evaluation using ROUGE metrics for text summarization and relevant metrics for multimedia content understanding showcases the effectiveness of our approach in producing clear and informative outputs. These advancements hold promising possibilities for information retrieval, content analysis, and user interaction across various applications.

List of abbreviations

Sr no.	Short form	Abbreviated form
1	LLM	Large Language Model
2	AI	Artificial Intelligence
3	NLP	Natural Language Processing
4	PDF	Portable Document Format
5	LSTM	Long Short-Term Memory
6	VAE	Variational Autoencoder
7.	RNN	Recurrent Neural Network
8	TXT	Plain Text
9	DOC	Document
10	PPT	Presentation
11	APIs	Application Programming Interface
12	ROUGE	Recall-Oriented Understudy for Gisting Evaluation
13	TTS	Text-to-Speech
14	GUI	Graphical User Interface
15	ROUGE-N	Recall-Oriented Understudy for Gisting Evaluation - N-gram.

List of figures

Chapter I: Introduction

1.1 Introduction

In today's information-overloaded digital environment, ReadWiz: AI Assistant for Mastering Any Content is your go-to resource. Whether you're a professional, student, or researcher, ReadWiz is designed to make learning easier for everyone. It provides strong tools for understanding complicated materials with ease. With its simple UI and innovative technology, ReadWiz makes it simple to extract important insights from lengthy papers and articles. It is a partner in the hunt for knowledge rather than just a tool. By leveraging advanced techniques in natural language processing (NLP) and large language models (LLM), ReadWiz adeptly handles multimedia content, rephrases intricate topics, and delivers concise summaries. ReadWiz is your reliable companion for subject mastery, always adapting to your demands.

1.2 Motivation

The motivation for developing ReadWiz: AI Assistant for Mastering Any Content comes from the need to address today's difficulties of too much information and making learning easy for everyone. We are bombarded with text in the modern world, and it can be challenging to stay along. Whether you are a professional, researcher, or student, we aim to make life easier for everyone. Our aim is to provide people with a tool that simplifies complicated knowledge. We think that everyone is able to acquire knowledge more easily when the material is clear and easy to grasp.

1.3 Problem Statement

The problem we aimed to solve with this project is the lack of accessible and comprehensive tools for managing the overwhelming volume and variety of digital content available today. We recognized that users face challenges in efficiently processing, understanding, and interacting with diverse types of information, including text, PDFs, videos, images, and other document formats. Existing systems often lack user-friendly interfaces and fail to offer a holistic set of features, such as text summarization, paraphrasing, and interactive capabilities with PDF files. Our goal was to develop a solution that addresses these shortcomings by providing intuitive interfaces and a comprehensive suite of functionalities to empower users to navigate and comprehend digital content more effectively.

1.4 Existing Systems

Several existing systems offer features akin to those provided by ReadWiz. SummarizeBot condenses lengthy texts into concise summaries, while Paraphrase Online aids in rewriting text to simplify language or avoid plagiarism. PDF Expert facilitates comprehensive PDF document management, and MindMeister enables visual organization of ideas through mind mapping. NaturalReader converts written text into spoken words for auditory consumption. However, ReadWiz stands out by integrating multiple functionalities such as text summarization, paraphrasing, PDF interaction with chat capabilities, mind mapping, text-to-speech, and image and video summary, along with summary analytics, into a single, user-friendly platform.

1.5 Lacuna of existing systems

Existing text processing systems frequently struggle to satisfy the wide range of user demands due to a number of issues. A major disadvantage is the absence of full integration across various functionalities. Although several systems do particularly well at certain tasks, such as paraphrase or text summarization, they frequently fall short of offering a full solution. Furthermore, a lot of current systems are not very user-friendly; in order to do certain activities, users frequently must switch between different platforms or tools. The restricted support for multimedia material processing, including picture and video summarization, is another flaw in these systems that limits their usefulness and accessibility. These drawbacks attract attention to the need for a more adaptable and user-friendly solution, such as ReadWiz, which solves these issues by combining a variety of features into a single platform.

1.6 Relevance of the project

In today's knowledge-centric world, when people and organizations struggle with information overload and advanced textual material, the project is extremely relevant. ReadWiz provides an extensive tool for understanding multimedia material, summarizing text, and paraphrasing, addressing the crucial need for effective information processing. It increases productivity and knowledge transfer by simplifying the process of obtaining important ideas from long materials. Additionally, ReadWiz encourages users from a variety of areas and backgrounds to have easy access to knowledge, enabling them to further their objectives and make well-informed judgments.

Chapter II: Literature survey

A: Textlytic: Automatic Project Report Summarization Using NLP Techniques. [1]

Authors: Menon, Riya & Tolani, Namrata & Tolamatti, Gauravi & Ahuja, Akansha & R L, Priya

The paper proposes a methodology for automatic project report summarization using NLP techniques. The approach involves preprocessing the document, splitting it into sections, and summarizing each section. The methodology uses PyPDF2 and Camelot libraries for document processing, word2vec for extractive summarization, and K-means clustering for grouping word vectors. The resulting summary is generated by combining the summaries and figures/diagrams of each section. The system also provides a user-friendly web application interface for convenience.

B: Hybrid Algorithm to Generate Summary of Documents by Extracting Keyword. [2]

Authors: Surbhi Patel, Devanshi Parikh, & Hiren Joshi

The paper presents a hybrid algorithm for generating summaries of documents by extracting keywords. The algorithm follows a step-by-step process that includes identifying the number of pages, performing N-gram analysis, removing stop words, applying stemming, and measuring word similarity using WordNet. The results are then displayed. The automatic keyword extraction enables the identification of key words, phrases, or segments that describe the meaning of the document. The proposed algorithm combines multiple methods and algorithms, such as Text Rank, RAKE, and TAKE, to achieve higher recall and f-measure compared to individual extractors.

C: A Deep Generative Framework for Paraphrase Generation. [3]

Authors: Ankush Gupta, Arvind Agrawal, Prawaan Singh, & Piyush Rai

The paper proposes a deep generative framework for paraphrase generation. It combines deep generative models (VAE) with sequence-to-sequence models (LSTM) to generate paraphrases given an input sentence. The models are trained using stochastic gradient descent with a fixed learning rate and dropout rate. The proposed method conditions both the encoder and decoder sides of the VAE on the original sentence to generate paraphrases. The proposed method outperforms state-of-the-art methods in terms of performance improvement, as demonstrated by quantitative evaluation on a benchmark paraphrase dataset. The generated paraphrases are well-formed, grammatically correct, and relevant to the input sentence, as indicated by qualitative human evaluation.

Chapter III: Requirement Gathering for the Proposed System

3.1 Introduction to requirement gathering

Requirement gathering is a crucial initial phase in the software development process, where the focus is on identifying and documenting the needs and expectations of stakeholders for a given project. This process involves gathering information about the desired features, functionalities, and constraints of the software system to be developed. The main objective of requirement gathering is to establish a clear understanding of what the software should accomplish and to define the scope of the project. This is achieved through various techniques such as interviews, surveys, workshops, and documentation analysis, where stakeholders, including clients, end-users, and domain experts, are actively engaged to elicit and prioritize their requirements.

3.2 Functional requirements

Functional requirements specify the specific functionalities and features that the system must provide to meet the needs of stakeholders and users. Functional requirements for our system are:

- Text Summarization: Accept input text of varying lengths and generate concise summaries of input text based on specified parameters (e.g., summary length).
- Paraphrasing: Accept input text or files for paraphrasing. Provide paraphrased versions of input text that preserve the original meaning. It gives 3 paraphrased outputs.
- Chat with PDF: Enable users to upload PDF documents for interactive chat. Extract text content from uploaded PDF files for chat interactions.
- Multimedia Content Summarization: Support the summarization of multimedia content, including videos and images.
- Text-to-Speech Functionality: It allows users to convert output texts into spoken audio.

3.3 Non functional requirements

Non-functional requirements define the quality attributes and constraints that the system must adhere to. In our project, non-functional requirements includes:

- Performance: The system should respond to user interactions within acceptable response times, even under peak loads. Text summarization, paraphrasing, and multimedia content processing should occur efficiently to minimize processing delays.

- Reliability: The system should be reliable and available for use at all times, with minimal downtime for maintenance or updates.
- Scalability: The system should be scalable to accommodate increasing numbers of users and growing volumes of data.
- Usability: Text summarization, paraphrasing, and multimedia content processing features should be accessible and straightforward to use, even for users with limited technical expertise.

3.4 Hardware, Software, Technology and tools utilized

- Hardware
  - Standard computing hardware including processors, memory, storage device
- Software
  - Python as primary programming language
- Technology
  - Tensorflow to implement neural networks
  - Google Gemini

3.5 Constraints

- Technological Constraints: The system must be compatible with specific natural language processing (NLP) frameworks or libraries for text analysis and processing. Integration with third-party APIs or services for text-to-speech functionality may impose constraints on the selection of compatible services and technologies.
- Resource Constraints: Limited computational resources, such as processing power and memory, may impact the system's ability to handle large volumes of data efficiently.
- Operational Constraints: Integration with existing systems or platforms, such as network bandwidth limitations or server capacity, may impact system performance and scalability, particularly during peak usage periods.
- User Constraints: User preferences and expectations regarding the user interface design, functionality, and accessibility may impose constraints on the system's usability and user experience. Multilingual or multicultural user bases may require support for diverse languages, dialects, or cultural norms, imposing constraints on content processing and interaction.
- Environmental Constraints: Environmental factors, such as internet connectivity issues or device compatibility limitations, may impact the accessibility and usability of the system for users in different geographical locations or environments.

Chapter IV: Proposed Design

4.1 Block diagram of system

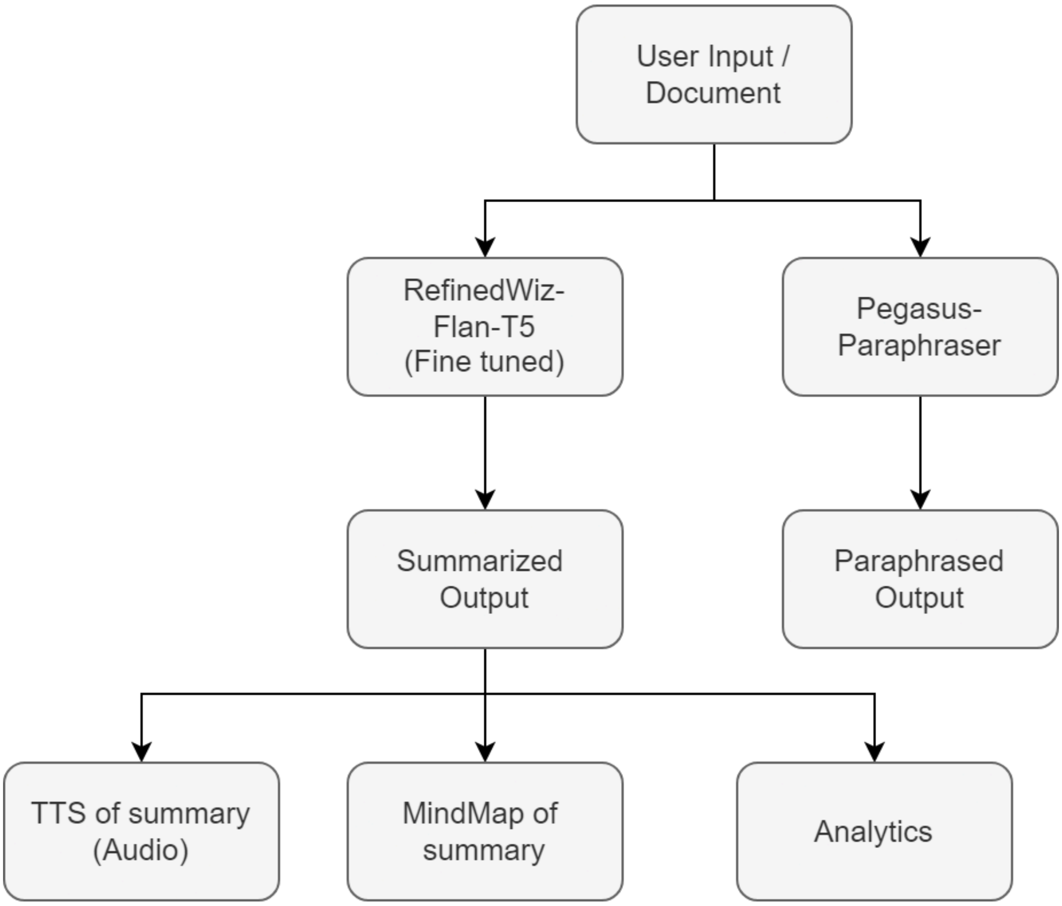

Figure 4.1 Block Diagram of system

1. User Input/Document: This section enables users to input text directly or upload documents in various formats such as Word, PDF, PPT, or TXT, providing flexibility in content input methods.

2. RefinedWiz (Fine-tuned Model): RefinedWiz represents the fine-tuned model responsible for processing the input text and generating a summary. It refines the input data and prepares it for further analysis and summarization.

3. Pegasus-Paraphraser: Powered by Pegasus, an advanced language model, the Paraphraser module rephrases the input text while maintaining its original meaning. This functionality is crucial for tasks like content rewriting, language learning, and avoiding plagiarism.

4. Summarized Output: After processing by RefinedWiz, this block produces a concise summary of the input text, highlighting key points and essential information extracted from the original content.

5. Paraphrased Output: Utilizing the capabilities of Pegasus, this block produces paraphrased versions of the input text. These paraphrases offer alternative expressions while preserving the core ideas, enhancing the accessibility and understanding of the content.

6. TTS (Text-to-Speech) of Summary (Audio): ReadWiz offers a Text-to-Speech feature that converts the summarized text into audio format. Users can listen to the summary instead of reading it, catering to different learning preferences and accessibility needs.

7. MindMap of Summary: To aid in visualizing the structure and key points of the summarized text, ReadWiz generates a mind map. This visual representation helps users grasp the relationships between different concepts and navigate the content more effectively.

8. Analytics: The Analytics block provides insights into the differences between the original text and the summary in terms of word count, character count, line count, and other relevant metrics, aiding users in evaluating the summarization effectiveness.

4.2 Project Scheduling & Tracking using Timeline / Gantt Chart

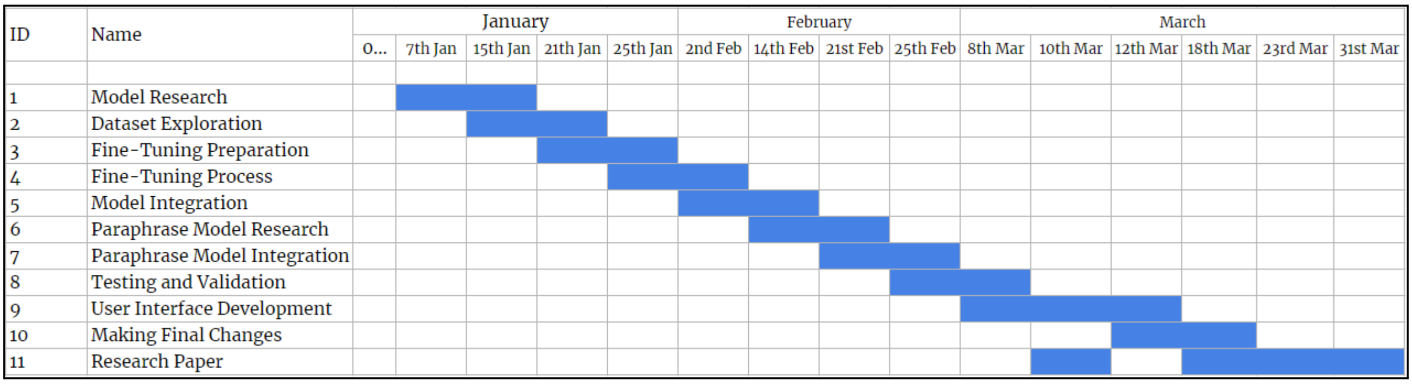

Figure 4.2 Gantt Chart

Chapter V: Implementation of Proposed System

5.1 Methodology employed for development

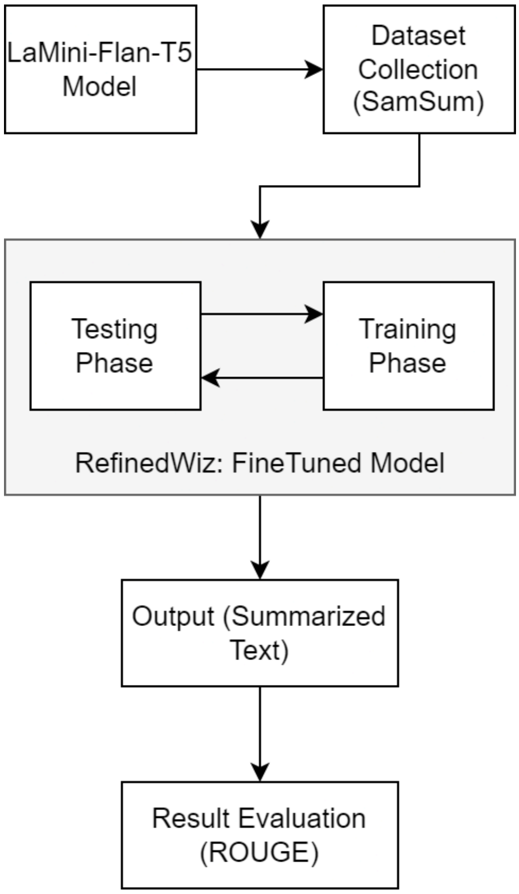

Figure 5.1 Methodology followed for Summarizer

- Model Selection and Preparation: We carefully choose the LaMini-Flan-T5 Model for its excellent text summarization capabilities. Before using it, we make sure the model is ready by preparing it and optimizing its settings to improve its performance.

- Dataset Collection: To train our summarization model effectively, we acquire the SAMSum dataset, containing over 16,000 chat-like conversations with accompanying summaries. This diverse dataset encompasses various linguistic styles and conversation types, providing valuable training data for our model.

- Fine-Tuning Model: The LaMini-Flan-T5 Model undergoes a fine-tuning process using the SAMSum dataset. Through this process, the model learns to better understand and summarize text, refining its abilities to generate accurate and concise summaries tailored to specific input text. Additionally, the fine-tuned model is referred to as "RefinedWiz," representing its enhanced capabilities in text summarization.

- Output Text Generation: Once the model is fine-tuned, it becomes capable of generating summaries based on input text. We focus on ensuring that these summaries are not only accurate but also coherent and relevant to the original text, making them easier for users to understand and extract key information from.

- Result Evaluation: To assess the effectiveness of our summarization model, we employ the ROUGE metric, which measures the similarity between the generated summaries and reference summaries. By analyzing ROUGE scores, we can evaluate the quality of our model's output and make any necessary adjustments to improve its performance.

5.2 Algorithms and flowcharts for the respective modules developed

Algorithm for Summarize Module:

- Input: Users can provide a document (PDF, TXT, DOC, PPT, Image) or text input or a video link, which is then preprocessed to remove any irrelevant information or noise.
- Process: The text is tokenized into sentences, allowing for a more granular analysis of the content. This step ensures that the summarization process captures the key points of each sentence.
- Action: The preprocessed text is fed into the RefinedWiz model, which has been fine-tuned to generate accurate and concise summaries. The model leverages advanced natural language processing techniques to identify the most important information and condense it into a summary.
- Output: The module produces a summarized version of the input text, providing users with a concise overview of the original content. Additionally, analytics are generated to compare the input data with the summary, including metrics such as word count, character count, and line count.

Algorithm for Paraphrase Module:

- Input: Similar to the summarization module, users input a document or text that they want to paraphrase for better understanding or to avoid plagiarism.
- Process: The input text undergoes preprocessing, which involves cleaning and formatting to ensure consistency and accuracy.
- Action: The module utilizes the Pegasus paraphrase model to generate multiple paraphrased versions of the input text. Pegasus is capable of understanding the context and semantics of the input text, allowing it to produce paraphrases that retain the original meaning while using different wording.
- Output: Three variations of the input text are generated, providing users with alternative formulations of the same content. This enables users to choose the paraphrase that best suits their needs or preferences.

Algorithm for Chat With PDF Module:

- Input: Users upload a PDF document containing the information they want to inquire about or discuss.

- Process: The module extracts the text content from the PDF document and preprocesses it to prepare it for analysis.
- Action: Users can ask questions related to the content of the PDF document, which are then processed by the system. The module employs the Gemini model, a chatbot specifically trained on the extracted text, to provide relevant and accurate answers to user queries.
- Output: Concise answers to user questions are displayed, allowing users to interact with the content of the PDF document in a conversational manner. This enhances user engagement and facilitates knowledge retrieval from the document.

5.3 Datasets source and utilization

Datasets Source:

- The SAMSum dataset, sourced from Hugging Face, comprises around 16,000 messenger-like conversations alongside summaries.
- Curated by linguists proficient in English, SAMSum conversations mimic real-life messenger exchanges, capturing various styles, registers, and content nuances.
- Annotated with summaries reflecting essential conversation points, SAMSum aims to provide concise third-person perspectives on the discussions.

Utilization in the Project:

- SAMSum plays a pivotal role in fine-tuning the text summarization model, particularly the RefinedWiz component.
- Fine-tuning entails training the model on a subset of SAMSum data to enhance its summarization capabilities.
- Leveraging the dataset's diverse conversational content, the fine-tuned model generates accurate and coherent summaries from input text.
- SAMSum also serves as a valuable resource for evaluating the model's performance and assessing its efficacy in capturing crucial information from conversations.

Chapter VI: Testing of Proposed System

6.1 Introduction to testing

Testing plays a crucial role in ensuring the system's reliability and functionality before it goes live. It involves systematically examining the system's performance under different scenarios to detect any issues, validate its functionality, and confirm that it meets the expected requirements. Through rigorous testing, we aim to identify and address any potential defects or problems, ensuring a smooth user experience and minimizing the likelihood of errors or malfunctions.

6.2 Types of tests Considered

- Unit Testing: Individual components are tested in isolation to verify their correctness and functionality.
- Integration Testing: Ensures that different components/modules work seamlessly together when integrated into the system.
- Functional Testing: Validates that the system's functionalities align with specified requirements and user expectations.
- Regression Testing: Verifies that recent changes in the codebase do not introduce unintended side effects or break existing functionalities.
- Performance Testing: Measures the system's responsiveness, stability, and scalability under various load conditions to ensure optimal performance.
- Security Testing: Identifies and addresses vulnerabilities to ensure the system's protection against potential security threats and breaches.
- Usability Testing: Assesses the system's ease of use and intuitiveness from an end-user perspective to enhance user experience.
- Compatibility Testing: Validates that the system functions correctly across different platforms, devices, and environments to ensure widespread accessibility and usability.

6.3 Various test case scenarios considered

- Summarization Accuracy: Ensure that the generated summaries accurately capture the key information from the original text across various document types and lengths.
- Paraphrasing Quality: Validate the quality of paraphrased texts by comparing them with the original input to ensure they preserve the original meaning while offering alternative expressions.
- PDF Chat Interaction: Test the chat functionality with PDF documents to ensure accurate extraction of information and relevant responses to user queries.

- Mind Map Visualization: Verify the effectiveness of the mind map feature in representing the summarized content's structure and key points for enhanced comprehension.
- Text-to-Speech (TTS) Conversion: Assess the clarity and coherence of the audio summaries generated through the TTS functionality to ensure easy comprehension for users.
- Image and Video Summary Accuracy: Evaluate the accuracy of descriptive summaries generated for multimedia content, ensuring they effectively convey the content's essence.
- Summary Analytics: Validate the accuracy of summary analytics, including metrics like word count, character count, and similarity scores between original and summarized texts.
- User Interface Interaction: Test the user interface for intuitiveness, responsiveness, and accessibility across different devices and screen sizes to ensure a seamless user experience.

6.4 Inference drawn from the test cases

The test cases conducted on ReadWiz revealed insights into its performance and functionality. Summarization accuracy varied based on text complexity, with shorter documents yielding more accurate summaries. Paraphrasing quality was satisfactory, preserving original meaning while offering diverse expressions. PDF chat interaction accurately extracted information and provided relevant responses. The mind map visualization aided in understanding content structure and key points. Text-to-speech conversion produced clear audio summaries, enhancing accessibility. Image and video summaries were generally accurate, with occasional discrepancies. Summary analytics provided valuable insights, aiding in content assessment. The user interface was intuitive, contributing to a seamless experience.

Chapter VII: Results and Discussion

7.1 Screenshots of User Interface (UI) for the respective module

Summarizer:

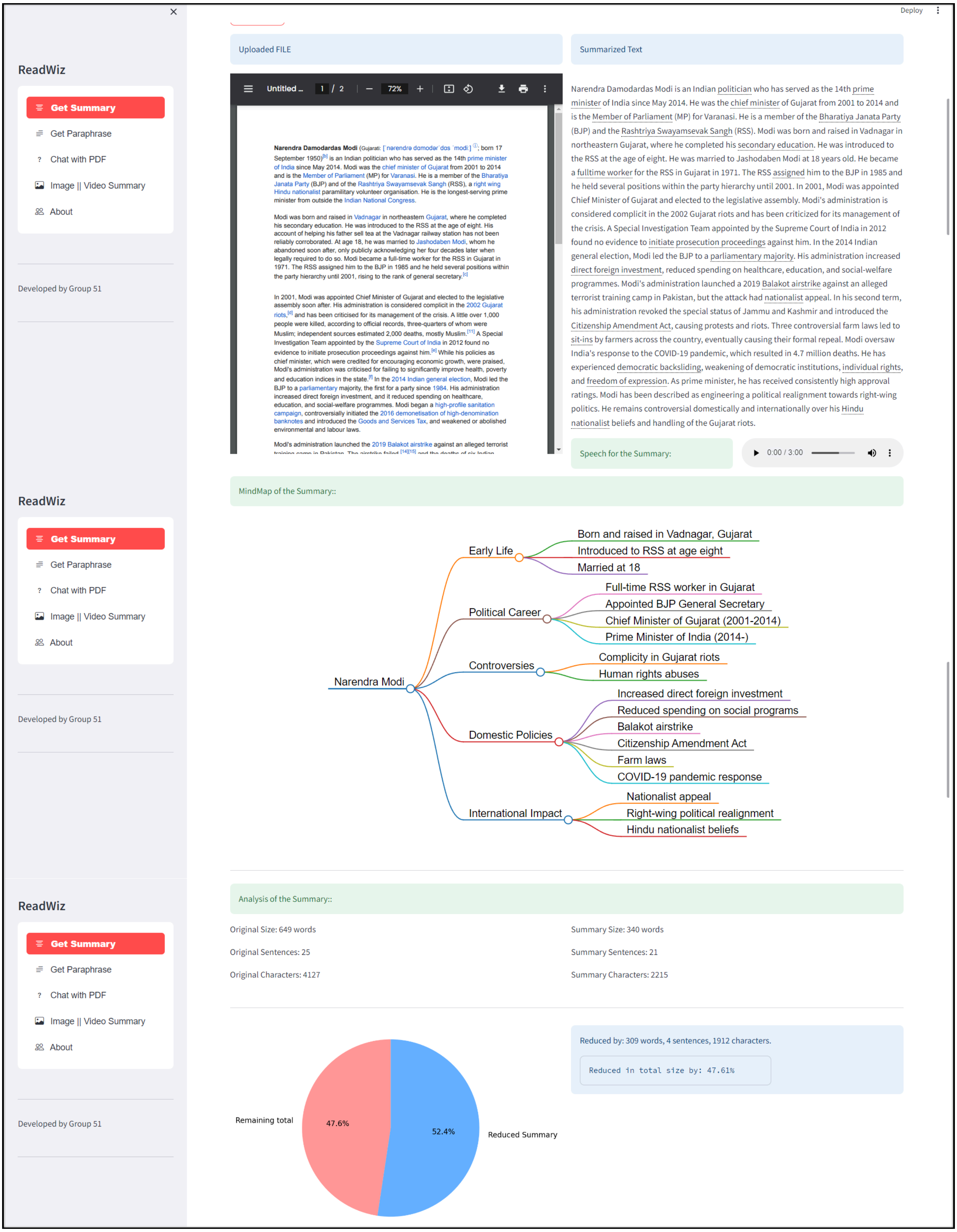

Figure 7.1.a Text Summarizer Module

Paraphraser:

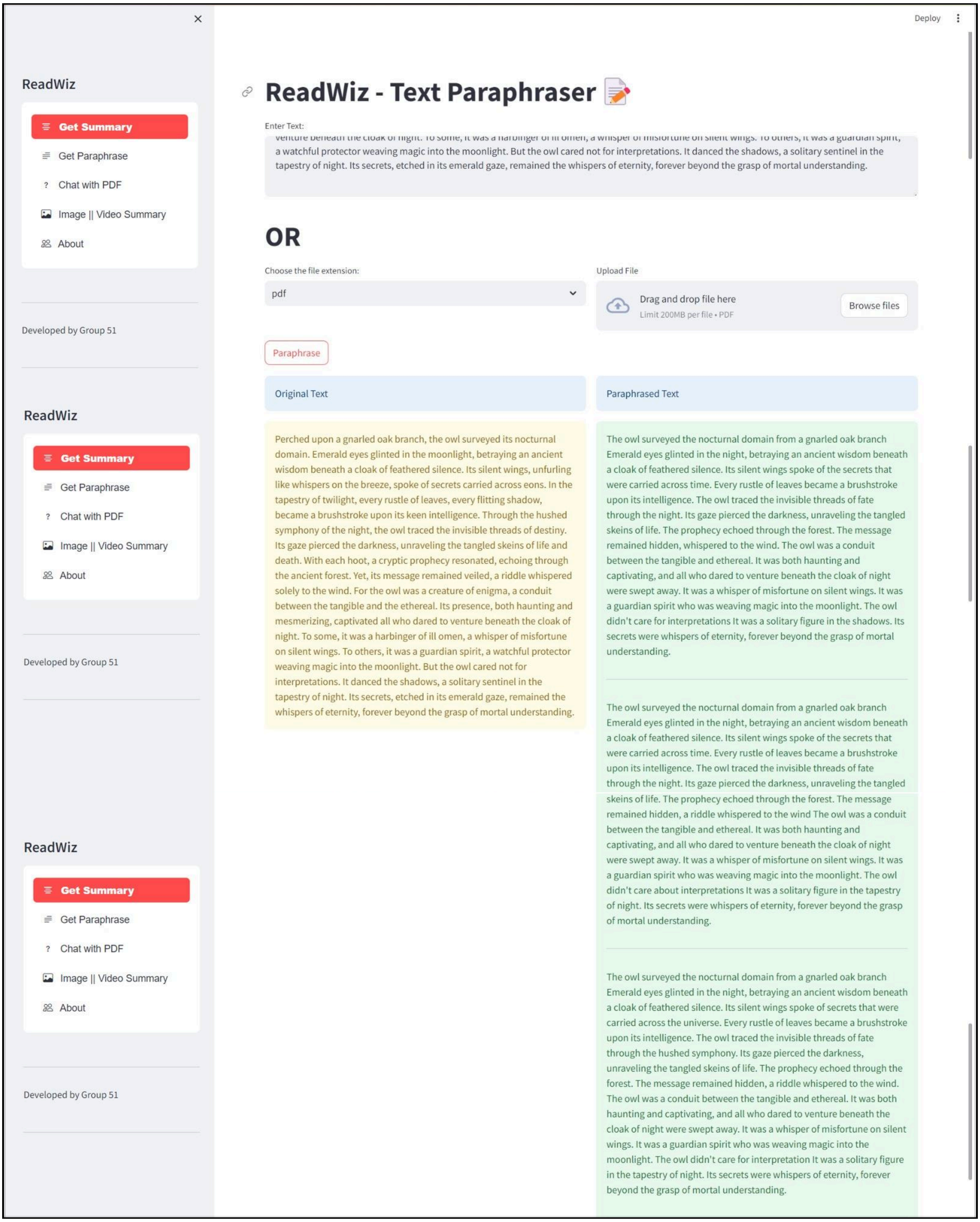

Figure 7.1.b Text Paraphraser Module

Chat with PDF:

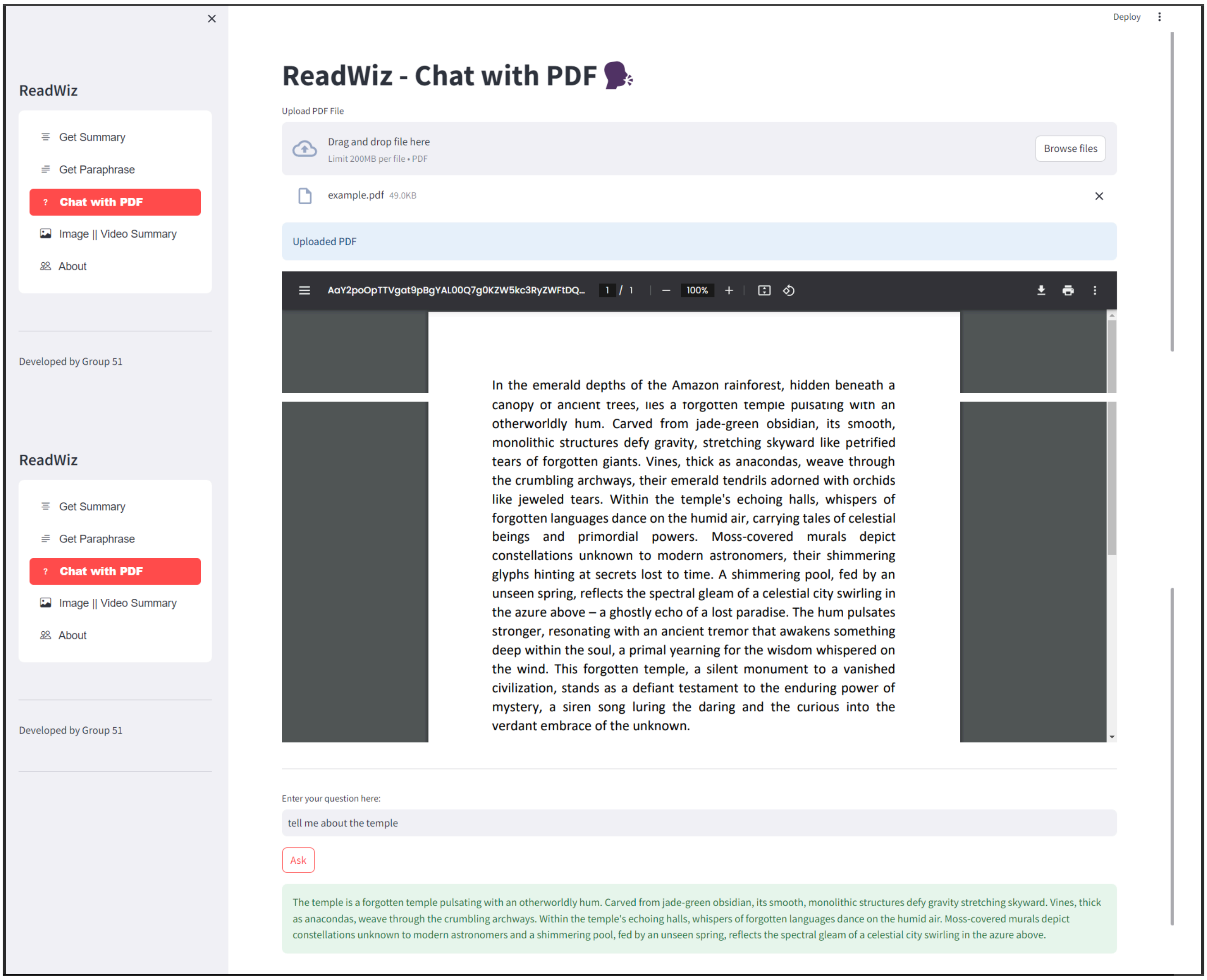

Figure 7.1.c Chat with PDF Module

Image and Video Summarizer:

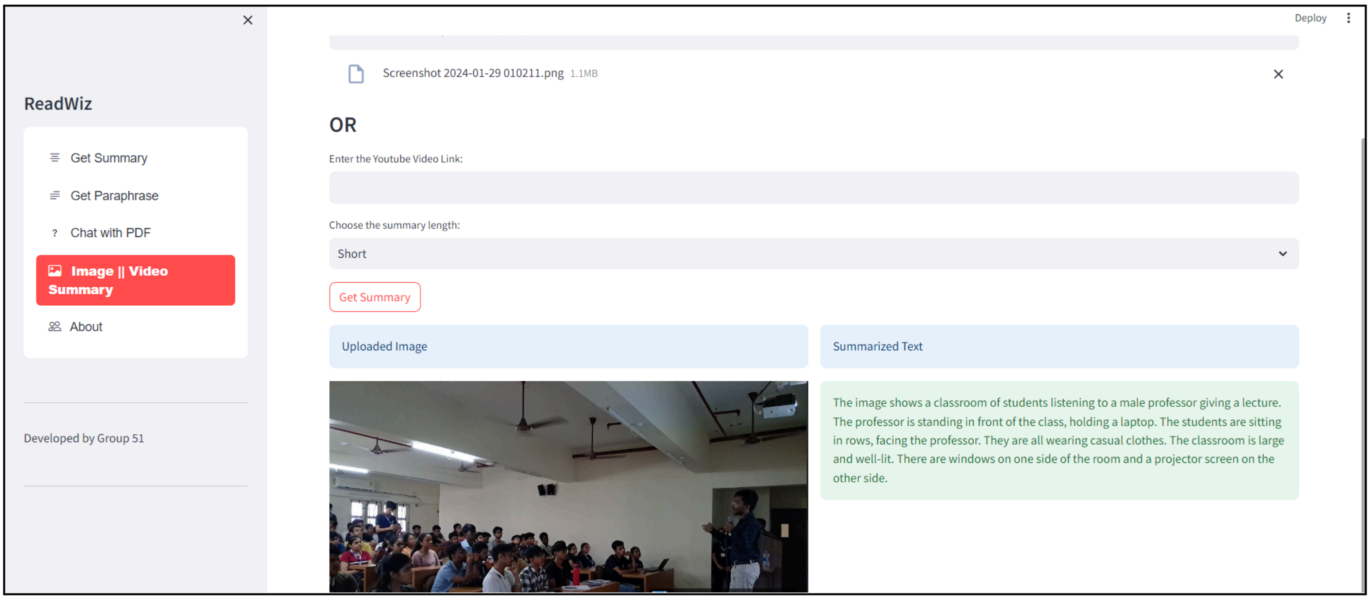

Figure 7.1.d Image and Video Summarizer Module (*Image Summary*)

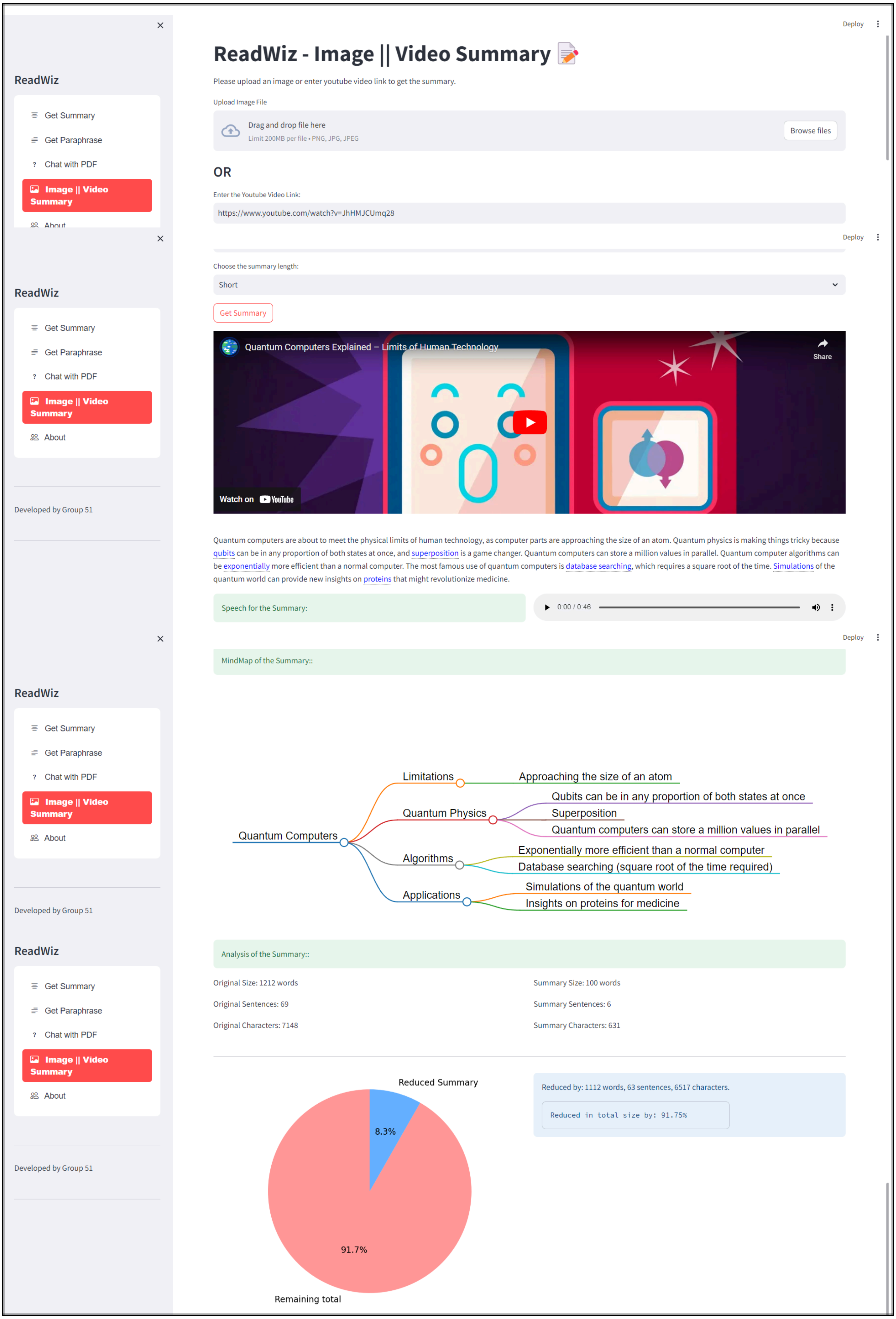

Figure 7.1.e Video Summarizer (*Youtube Video Summary*)

7.2 Performance Evaluation measures

For evaluating the performance of ReadWiz Summarizer Module, the ROUGE metric was employed as the primary evaluation measure. ROUGE assesses the quality of generated summaries by comparing them with reference summaries, measuring factors like overlap in n-grams, word sequences, and sentence similarity. This metric provides valuable insights into the coherence, relevance, and informativeness of the generated summaries. Additionally, other performance evaluation measures, such as precision, recall, and F1-score, may also be utilized to assess the system's summarization accuracy and effectiveness.

- Precision: Precision measures the accuracy of the system's generated summaries by calculating the ratio of relevant information presented in the generated summary to the total information included in the summary.
- Recall: Recall evaluates the comprehensiveness of the system's generated summaries by determining the ratio of relevant information covered in the generated summary to the total relevant information available in the source document.
- F1 Score: The F1 score provides a balance between precision and recall, calculated as the harmonic mean of precision and recall. It offers a single metric to assess the overall effectiveness of the system's summarization performance.

Example:

LLM Model	ROUGE-1			ROUGE-2			ROUGE-L		
	R	P	F	R	P	F	R	P	F
RefinedWiz-Flan-T5-248M (Fine tuned)	49.18	38.02	41.29	17.19	12.14	13.25	45.91	34.72	38.11

Figure 7.2 Performance of RefinedWiz model on sample data

7.3 Input Parameters / Features considered

For evaluation based on ROUGE metrics, we considered the following input parameters:

- ROUGE-N Precision: Measures the proportion of overlapping n-grams (contiguous sequences of n words) between the generated summary and the reference summary.
- ROUGE-N Recall: Calculates the proportion of overlapping n-grams between the generated summary and the reference summary relative to the reference summary.
- ROUGE-N F1 Score: Represents the harmonic mean of precision and recall, providing a balanced measure of summarization quality.
- ROUGE-L Precision: Evaluates the precision of the longest common subsequence (LCS) between the generated summary and the reference summary.

- ROUGE-L Recall: Assesses the recall of the LCS between the generated summary and the reference summary relative to the reference summary.
- ROUGE-L F1 Score: Computes the F1 score of the LCS, indicating the overall effectiveness of the summarization process.

7.4 Graphical and statistical output

Outputs for Summarizer:

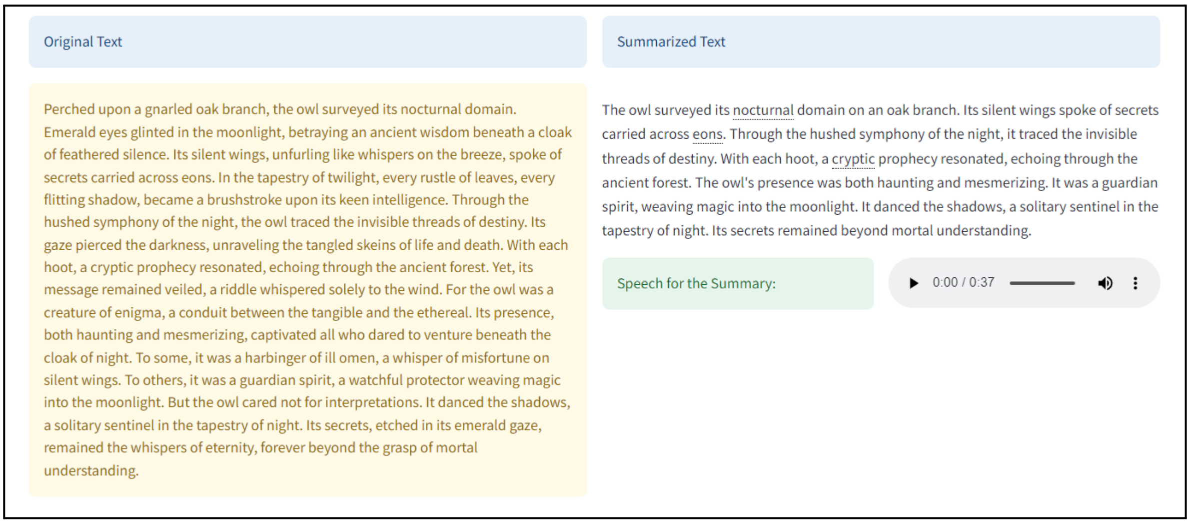

Figure 7.4.a Generating summary of the sample paragraph

Following the creation of the summary according to the input text, we moved our attention to evaluating the performance of the summarizing process. We evaluated the quality of the produced summaries using recognized assessment measures by utilizing a thorough testing technique. Using Rouge testing, a popular approach to assess text summarizing algorithms' performance, the input text was carefully analyzed. The findings were then carefully examined and graphically shown to give an in-depth understanding of the effectiveness of the summarizing procedure. This included creating a line graph using the Rouge scores that were obtained during the testing phase to provide an understandable and straightforward representation of the summarizing model's performance. Through this careful review and visualization method, we wanted to give consumers significant insights into the summary process and its outcomes, facilitating informed decision-making and enhancing overall user experience.

Rouge results for this generated Summary:

{'rouge-1':
{'r': 0.5254237288135594, 'p': 0.47692307692307695, 'f': 0.4999999950117067}},
{'rouge-2':
{'r': 0.2898550724637681, 'p': 0.25, 'f': 0.26845637086617773}},
{'rouge-l':
{'r': 0.4745762711864407, 'p': 0.4307692307692308, 'f': 0.45161289823751305}}

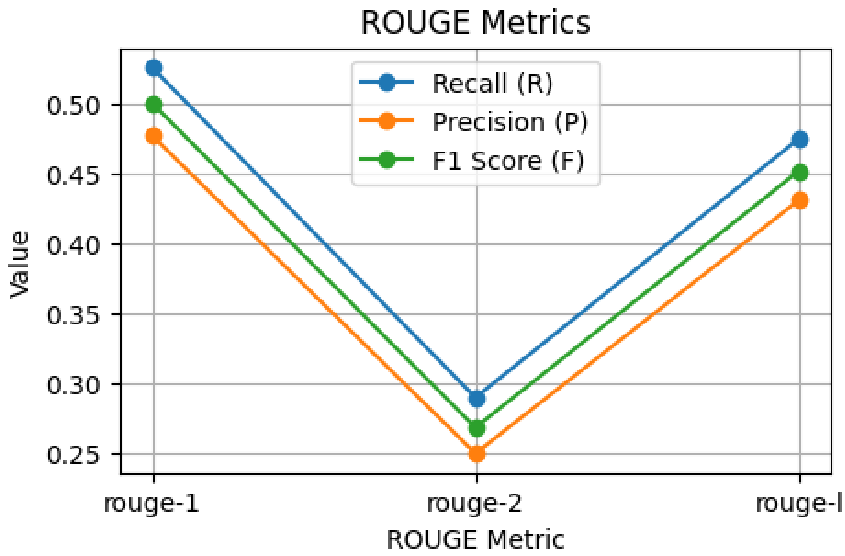

Figure 7.4.b ROUGE metrics of RefinedWiz model on sample data

7.5 Comparison of results with existing systems

In our comparative analysis for the Summarizer task between the RefinedWiz model and an online tool, we aimed to evaluate their performance in text summarization based on various ROUGE metrics. The comparison was conducted on a subset of rows extracted from the SAMSum dataset, with the average scores considered for each metric.

The RefinedWiz model achieved higher scores across almost all ROUGE metrics compared to the online tool, indicating its superior performance in text summarization.

LLM Model	ROUGE-1			ROUGE-2			ROUGE-L		
	R	P	F	R	P	F	R	P	F
RefinedWiz Model	53.21	33.71	40.48	16.95	08.80	11.00	50.54	31.85	38.32
Online Tool	51.63	31.46	38.51	15.54	09.36	11.57	47.15	29.05	35.42

Figure 7.5 Comparison with Online Tool

For ROUGE-1, which measures the overlap of unigram units between the generated summary and the reference summaries, the RefinedWiz model achieved a Recall of 53.21%, Precision of 33.71%, and F1-score of 40.48%. In comparison, the online tool scored slightly lower with a Recall of 51.63%, Precision of 31.46%, and F1-score of 38.51%.

Similarly, for ROUGE-2, which considers the overlap of bigram units, the RefinedWiz model outperformed the online tool with a Recall of 16.95%, Precision of 8.80%, and F1-score of 11.00%, compared to the online tool's Recall of 15.54%, Precision of 9.36%, and F1-score of 11.57%.

Finally, for ROUGE-L, which computes the longest common subsequence of words, the RefinedWiz model demonstrated superior performance with a Recall of 50.54%, Precision of 31.85%, and F1-score of 38.32%, while the online tool achieved a Recall of 47.15%, Precision of 29.05%, and F1-score of 35.42%.

Sample Example:

Input: *Olivia: Who are you voting for in this election? Oliver: Liberals as always. Olivia: Me too!! Oliver: Great.*

Actual Summary: *Olivia and Olivier are voting for liberals in this election.*

Online tool: *Olivia inquires about her voting preferences in the upcoming election, and Oliver confirms that she is a Liberal, which Olivia agrees with.*

RefinedWiz Summary: *Olivia and Oliver are voting for the Liberals in the election. Oliver is also a Liberal. They both agree that it's a great election.*

7.6 Inference drawn

From the comparison with ROUGE measures, it is clear that the RefinedWiz model performs better than the current online tool in several evaluation categories. RefinedWiz's better recall, precision, and F1 ratings indicate how well it summarizes textual material. This improved performance demonstrates how strong RefinedWiz is at producing concise and appropriate summaries from input texts or documents.

Through improved summarization accuracy, RefinedWiz helps users quickly extract important details and insights from long texts. As a result, RefinedWiz is unique as a useful tool for a range of uses, such as data processing jobs, organizing content, and educational research. Furthermore, these results underscore the need of utilizing cutting-edge natural language processing methods, like those utilized in RefinedWiz, to meet the increasing need for efficient text summarization solutions.

Chapter VIII: Conclusion

8.1 Limitations

While ReadWiz offers remarkable capabilities in text summarization, it is not without its limitations. Firstly, due to the complexity of processing large volumes of text, the summarization process may take some time to complete, especially for lengthy documents. Additionally, it's important to recognize that the generated summary is computer-generated and may not always capture the nuanced details of the original text with absolute perfection. Furthermore, when summarizing videos from platforms like YouTube, access to a transcript is essential, as ReadWiz relies on textual content for analysis and summarization. Moreover, utilizing the app efficiently may require a good computer specification to handle the processing demands effectively. Additionally, while the chat feature with PDF documents enhances interaction and accessibility, it may not always provide additional insights beyond the content already available in the document. Despite these limitations, ReadWiz remains a valuable tool for simplifying and enhancing the comprehension of textual material, offering users an efficient means to navigate through vast amounts of information.

8.2 Conclusion

In conclusion, our project aims to provide a simple and effective solution for summarizing text, paraphrasing content, and managing multimedia resources. We're committed to ensuring that our tool is user-friendly and accessible to everyone. By incorporating user feedback and continuously enhancing our system, we're confident that we can deliver a valuable resource for individuals seeking to streamline their digital content experience. We're excited about the possibilities our project holds for improving information accessibility and usability for users across various domains.

8.3 Future Scope

- Better GUI: Enhance the user interface with intuitive design elements, clear navigation, and customizable layouts to improve user experience and engagement.
- Multi-languages: Support multiple languages to cater to a diverse user base, allowing users to interact with the system in their preferred language for improved accessibility and inclusivity.
- Faster Processing: Optimize algorithms and system architecture to ensure faster processing times, enabling quick generation of summaries, paraphrases, and multimedia representations for enhanced efficiency and productivity.

- Comic/Video Representation: Transform text summaries into engaging comic strips or animated videos, adding visual and audio elements to make the content more engaging and memorable for users.
- Offline Access: Provide offline access to summarized content and functionalities, enabling users to access and interact with the system even without an internet connection for increased flexibility and convenience.
- Voice Control: Implement voice control features, allowing users to navigate, interact, and generate content using voice commands for hands-free operation and accessibility.

References

[1] Menon, Riya & Tolani, Namrata & Tolamatti, Gauravi & Ahuja, Akansha & R L, Priya. (2022). “Textlytic: Automatic Project Report Summarization Using NLP Techniques.” 10.1007/978-981-16-7088-6_10.

[2] A. K. Mohammad Masum, S. Abujar, M. A. Islam Talukder, A. K. M. S. Azad Rabby and S. A.Hossain, "Abstractive method of text summarization with sequence to sequence RNNs," 2019 10th International Conference on Computing, Communication and Networking Technologies(ICCCNT), Kanpur, India, 2019, pp. 1-5, doi: 10.1109/ICCCNT45670.2019.8944620.

[3] Devanshi Parikh , Surbhi Patel , Dr. Hiren Joshi, 2020, Hybrid Algorithm to Generate Summary of Documents by Extracting Keywords, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 09, Issue 04 (April 2020)

[4] Edward, Shirly. (2018). Text-to-Speech Device for Visually Impaired People. International Journal of Pure and Applied Mathematics. 119.

[5] Zhang, Tianyi, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. "Benchmarking large language models for news summarization." arXiv preprint arXiv:2301.13848 (2023).

[6] S. Modi and R. Oza, "Review on Abstractive Text Summarization Techniques (ATST) for single and multi documents," 2018 International Conference on Computing, Power and Communication Technologies (GUCON), Greater Noida, India, 2018

[7] A. R. Pal and D. Saha, "An approach to automatic text summarization using WordNet," 2014 IEEE International Advance Computing Conference (IACC), Gurgaon, India, 2014

Appendix

Research paper details

Attached separately

Review I sheet:

Industry / Inhouse:
Research / Innovation: **Project Evaluation Sheet 2023-24** **Class: D12** B

Title of Project (Group no): ReadWiz (Text Summarizer and Paraphraser) 51

Group Members: Nikhil Dhanwani (12), Chirag Santwani (50), Manraj Singh Virdi (66)

	Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (3)	Applied Engg & Mgmt principles (3)	Life - long learning (3)	Professional Skills (5)	Innovative Approach (5)	**Total Marks (50)**
Review of Project Stage 1	4	4	4	3	4	2	2	2	2	3	3	3	5	4	45

Comments:

Dr. Sujata Kadam
Name & Signature Reviewer1

	Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (3)	Applied Engg & Mgmt principles (3)	Life - long learning (3)	Professional Skills (5)	Innovative Approach (5)	**Total Marks (50)**
Review of Project Stage 1	4	4	4	3	4	2	2	2	2	3	3	3	5	4	45

Comments:

Date: 10th February, 2024

Manisha Mathur 10/2/24
Name & Signature Reviewer2

Review II sheet:

Inhouse/ Industry _Innovation/Research:

Sustainable Goal:

Project Evaluation Sheet 2023 - 24

Class: D12 A/B/C

Group No.: 51

Title of Project:
ReadWiz (The all-in-one Text Processing tool)

Group Members: Nikhil Dhanwani (12), Chirag Santwani (50), Manraj Singh Virdi (66)

Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (2)	Applied Engg&Mgmt principles (3)	Life - long learning (3)	Professional Skills (3)	Innovative Approach (3)	**Research Paper (5)**	**Total Marks (50)**
4	4	5	2	5	2	2	2	2	2	3	3	3	3	3	45

Comments: Good work

Dr. Sujata Kadam
Name & Signature Reviewer1

Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (2)	Applied Engg&Mgmt principles (3)	Life - long learning (3)	Professional Skills (3)	Innovative Approach (3)	**Research Paper (5)**	**Total Marks (50)**
4	4	5	2	5	2	2	2	2	2	3	3	3	3	3	45

Comments:

Date: 9th March, 2024

Manisha Mathur
Name & Signature Reviewer 2

Analysis of Transformer based Pre-trained Summarization Models

Prof. Manisha Mathur
Assistant Professor, Computer Engineering
Vivekanand Education Society's Institute of Technology
Mumbai, India
manisha.mathur@ves.ac.in

Manraj Singh Virdi
Student, Computer Engineering
Vivekanand Education Society's Institute of Technology
Mumbai, India
d2021.manrajsingh.virdi@ves.ac.in

Nikhil Dhanwani
Student, Computer Engineering
Vivekanand Education Society's Institute of Technology
Mumbai, India
d2021.nikhil.dhanwani@ves.ac.in

Chirag Santwani
Student, Computer Engineering
Vivekanand Education Society's Institute of Technology
Mumbai, India
d2021.chirag.santwani@ves.ac.in

***Abstract*—In order to efficiently retrieve and comprehend large volumes of information, automated text summarization plays a crucial role in compressing the material into brief and understandable summaries. In this study, a thorough comparison of six innovative transformer-based summarization models sourced from the Hugging Face model hub is presented. Using ROUGE scores to evaluate sample outputs in both quantitative and qualitative terms, each model's performance and capabilities are carefully assessed, clarifying its unique advantages, disadvantages, and complexities. The results reveal significant variations in the models' summarization quality and effectiveness, likely stemming from differences in their architectures, pre-training sets, and fine-tuning techniques. The objective is to equip both students and professionals with the necessary knowledge to make informed decisions when selecting models for various summarization tasks, thereby laying the groundwork for future developments in transformer-based text summarization techniques.**

***Index Terms*—Automated text summarization, Transformer-based, ROUGE scores, Model performance, Pre-training sets, Fine-tuning techniques, Text summarization techniques.**

I. Introduction

In today's digital world, we can access a huge quantity of information with a few touches on our devices. This need more effective methods of understanding and managing all of this data. Text summary helps by carefully arranging and dividing enormous amounts of text into brief summaries while preserving crucial facts. Summarized text is even more crucial in addressing the information overload due to the necessity for rapid and efficient methods of finding information. Additionally, by reducing time and effort on information processing duties, text summarizing contributes to increased productivity. By giving brief summaries, it enables users to concentrate on key elements without getting distracted by information.

The field of deep learning and natural language processing has developed so rapidly that models that have been trained are now freely accessible and helpful for many NLP applications, especially text summarization. However, with so many models available—from BERT to T5—selecting the optimal pre-trained model for a specific job might be challenging. In addition, while pre-trained models offer an excellent place to begin, they can be even more effective and adaptable by modifying them with data specific to a certain area.

The motive of this exploration is to perform a thorough evaluation of six innovative text summarization models sourced from the Hugging Face model hub [1], considering the challenges posed by information overload. Specifically, the models examined include Facebook's BART-Large-CNN [2], Falconsai's Text-Summarization [4], MBZUAI's LaMini-Flan-T5-248M [5], sshleifer's DistilBART-CNN-12-6 [7], MBZUAI's LaMini-T5-61M [8], and Tuner007's Pegasus_summarizer [9]. This evaluation aims to highlight the advantages, disadvantages, and potential of each model, utilizing recognized metrics like ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores [10] and detailed examination of sample outcomes. Additionally, the investigation will explore the effectiveness of fine-tuning techniques in enhancing these models' performance on text summarization tasks. Through this comparative analysis, valuable insights can be provided to professionals and scholars, aiding them in navigating the landscape of summarized text models and optimizing their utility for managing information overload.

II. Background

Text summarization stands as a fundamental task within the domain of Natural Language Processing (NLP), addressing the challenge of condensing extensive volumes of textual data into succinct and coherent summaries. With an exponential surge in the creation of research papers, news articles, social media posts, and various digital materials, the need for effective summarization methodologies becomes increasingly paramount. Beyond mere condensation, text summarization serves as a facilitator for diverse tasks such as document classification, outcome ranking, and tailored content suggestion.

A. Approaches to Text Summarization

There are two primary approaches for text summary in natural language processing: extractive and abstractive summarization. In extractive summarizing [13], the precise wording and organization of the original information is maintained by picking out significant lines or paragraphs from the text as a whole and then presenting them in the form of a summary. In order to pick the most useful words based on factors including relevance, significance, and repetition, this method uses algorithms. Abstractive summary [13], on the other hand, involves developing new content that clearly and logically communicates the key points of the original text. Abstractive approaches clarify and paraphrase text using natural language comprehension and production processes, which frequently provide summaries that are more concise and logical.

B. Large Language Models (LLMs)

Among Large Language Models (LLMs) [14], Transformer-based models such as BART, GPT and T5 have transformed text summarization by their capacity to recognize complex language patterns and meaningful connections. This is because they have been pre-trained on huge amounts of text data. These models show exceptional understanding and generation skills for human-like writing, which makes them useful for a broad series of jobs with respect to natural language processing, including summarization. By using the encoded knowledge stored inside its boundaries, LLMs are able to efficiently compress complicated documents while maintaining crucial information and context. Analyzing automatically generated summaries is still difficult, though. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores [12] are frequently utilized metrics; however, they have several drawbacks, like their insensitivity to semantic similarity and their incapacity to capture readability and overall consistency. Thus, human assessment continues to be necessary for gaining an in-depth knowledge of summary quality, completing automated measures such as ROUGE and opens the door for future advances in the field.

C. Fine Tuning Models

In addition, optimizing Large Language Models (LLMs) on datasets customized to certain tasks offers an appealing method to improve their summarization performance. By significantly altering the pre-trained model's parameters, fine-tuning allows it to adapt its enormous prior knowledge to the specifics of a given summarization job or topic. It has been shown that this procedure significantly improves performance, helping the model to provide summaries that are more precise, relevant, and consistent. A crucial topic of investigation in this work is the possibility of fine-tuning LLMs for summarization tasks, especially since the goal is to expand the boundaries of the field by optimizing model performance on particular datasets.

III. METHODOLOGY

In this section, the thorough methods used to assess and compare the efficacy of six cutting-edge text summarization algorithms are described. The approach aims to evaluate how well each model generates concise and cohesive summaries in various domains. The criteria for model selection are outlined, along with the datasets utilized for evaluation and any fine-tuning processes. The evaluation metrics adopted to measure summarization quality are specified. To enhance flexibility and consistency of the results, the experimental environment, including system data and computational capabilities, is also provided. This technique attempts to investigate the possible advantages of fine-tuning to enhance model performance along with highlighting the relative benefits and drawbacks of each model. A significant contribution to the field of automated text summarizing is sought through the implementation of a systematic and rigorous evaluation approach that will direct further studies and their application innovation.

A. Model Selection

Text summarization tools capabilities and performance are significantly affected by their model architecture. Among the many different designs that are accessible, BART (Bidirectional and Auto-Regressive Transformers) [16] a transformer-based model noted for being flexible to natural language processing applications. It can efficiently collect context and produce coherent text because to its bidirectional architecture and autoregressive decoding. T5 (Text-To-Text Transfer Transformer) [17], designed by Google AI, uses a unified architecture in which the inputs and outputs are expressed as text strings, making it suitable for a variety of NLP applications. FLAN T5 [18] a version of T5 that has been specially optimized for text summary, providing greater efficiency in producing concise and insightful summaries. Another Google AI model, Pegasus [19], is well known for its abstractive summarization skills. It can efficiently extract important information from source texts and reword or rephrase it. Several models utilize these architectures, including 'facebook/bart-large-cnn' based on BART, 'MBZUAI/LaMini-T5-61M' based on T5, 'MBZUAI/LaMini-Flan-T5-248M' based on FLAN T5, and 'tuner007/pegasus_summarizer' based on Pegasus.

Selected Models

- **Model 1: facebook/bart-large-cnn**
  This model, as cited in [3], created by the research team at Facebook AI, is an effective tool for text summarization. It is based on the robust design of BART (Bidirectional and Auto-Regressive Transformers). The architecture allows the model to understand complicated language connections and structures observed in textual data. Designed specifically for news item overviews, 'facebook/BART-Large-CNN' is excellent at collecting important texts, which reduces the amount of information that is consumed. By carefully tuning with the CNN Daily Mail dataset—a comprehensive collection of text-summary pairs—the algorithm achieves outstanding results on summarizing tasks.
- **Model 2: Falconsai/text_summarization**

Designed by Falcon AI [4], it is aimed at the medical text summarizing industry. Based on a foundation pre-trained on a wide range of medical literature, this innovative approach makes use of an optimized version of the T5 Large transformer model. The model is excellent at producing concise and coherent summaries that are appropriate to the complexities of the medical area because it focuses on medical texts, such as research papers, clinical notes, and documents. By applying a thorough method of fine-tuning on a composite dataset gathered from various fields, the 'falconsai/text_summarization' model ensures its versatility and efficacy in a wide range of tasks.

- **Model 3: MBZUAI/LaMini-Flan-T5-248M**
  A refined variant of Google's Flan-T5-base model, called 'MBZUAI/LaMini-Flan-T5-248M' [6], is optimized to work with the LaMini-instruction dataset, which contains about 2.58 million instruction samples. It provides compact, effective language models with a range of sizes, checkpoints, and structures as part of the LaMini-LM series. In contrast to other models in the series, the 'MBZUAI/LaMini-Flan-T5-248M' utilizes an encoder-decoder architecture with 248 million parameters, that is based on the T5 model. Its adaptability and usability in a variety of environments are guaranteed by its rigorous evaluations across several NLP tasks and human assessments, demonstrating its flexibility to a wide range of natural language processing activities, including text generation.
- **Model 4: sshleifer/distilbart-cnn-12-6**
  The 'sshleifer DistilBART CNN 12-6' [7] model is a pre-trained language model designed for text summarization tasks based on the CNN/DailyMail dataset. It is a compressed and simplified variation of the BART model. Using a distillation process, DistilBART transfers the knowledge and performance of the bigger BART model by teaching a smaller student version to behave like it. Specifically designed for text summarization, the 'distilbart-cnn-12-6' variation uses a Convolutional Neural Network (CNN) encoder architecture with 12 layers of encoder and 6 decoder levels. This architecture improves the model's summarizing abilities by assisting in the acquisition of local contextual information in the input text.
- **Model 5: MBZUAI/LaMini-T5-61M**
  A key part of the LaMini-LM series, the 'MBZUAI/LaMini-T5-61M' [8] model was carefully developed by the Mohamed bin Zayed University of Artificial Intelligence (MBZUAI). Optimized for instruction fine-tuning, this version of the T5-small model has been developed on the LaMini-instruction dataset, an extensive collection of 2.58 million samples. The model has 61 million parameters, balances both efficiency and performance, which makes it a great choice for a range of natural language processing applications. For this study, the model chosen is 'MBZUAI/LaMini-T5-61M' because of its stable structure, tailored instruction using domain-specific data, and alignment with the goals stated in the research report "LaMini-LM: A Diverse Herd of Distilled Models from Large-Scale Instructions." [6]. This model is a good fit for summarization experiments because it provides a strong combination of computational effectiveness and task-specific expertise.
- **Model 6: tuner007/pegasus_summarizer**
  The pegasus_summarizer [9] is a text summarizing model that is available to the public and has been developed to maximize the potential of Google AI's Pegasus architecture. Since Pegasus serves as its foundation, it undoubtedly specializes at abstractive summarization and is skilled at utilizing various words and phrasing to express the main idea of the source material. A refined version of Pegasus, the 'tuner007/pegasus_summarizer' model is designed to improve performance on text summarizing work. It seeks to further improve summarization abilities by optimizing the original Pegasus model, giving consistency and informativeness in produced summaries.

TABLE I
OVERVIEW OF TEXT SUMMARIZATION MODELS

Model Name	Description	Architecture	Size
facebook/bart-large-cnn	BART with CNN encoder, fine-tuned on CNN Daily Mail dataset	Encoder-Decoder (BART) with CNN encoder	406M params
Falconsai/ text_summarization	Fine-tuned T5 Large for medical text summarization	Transformer (T5)	60.5M params
MBZUAI/LaMini-Flan-T5-248M	Fine-tuned Flan-T5-based on Lamini-instruction dataset	Transformer (Flan-T5)	248M params
sshleifer/distilbart-cnn-12-6	Distilled BART with CNN encoder for efficient summarization tasks	Encoder-Decoder (Distilled BART) with CNN encoder	306M params
MBZUAI/LaMini-T5-61M	Compact T5 variant for summarization	Transformer (T5)	61M params
tuner007/ pegasus_summarizer	An optimized model based on the powerful Pegasus architecture	Pegasus-based model	-

Table I provides a summary of the models examined in this study. After providing an overview of each model, the process of preparing the data is delved into, describing the dataset used for assessment and any processing measures taken to ensure consistency and precision in the evaluation of the model's performance.

B. Dataset

SAMSum Dataset: The SAMSum dataset [20] includes around 16,000 messenger-like conversations with summaries for each. These dialogues were created by English-speaking linguists who were given the task to create interactions that matched their regular talks. The goal of the dataset is to replicate the subject matter distribution seen in actual messenger chats. Discussions display a broad range of records and styles, including semi-formal, formal, and casual conversations. The conversations were then annotated with brief summaries that were meant to capture the key points addressed in the dialogues. The Samsung R&D Institute Poland created the SAMSum dataset, which is licensed under a non-commercial agreement (CC BY-NC-ND 4.0) for use in research.

Data Splits:

- Train: 14,732 conversations
- Validation: 818 conversations
- Test: 819 conversations

Sample Instance:

- ”id”: ”13818513”,
  ”dialogue”:
  - ”Amanda: I baked cookies. Do you want some? Jerry: Sure! Amanda: I’ll bring you some tomorrow.”

  ”summary”:
  - ”Amanda baked biscuits and will bring Jerry some hereafter”,

C. Evaluation Strategy

In the research conducted, the main criteria utilized to evaluate the level of quality of the produced summaries was ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores [11]. The ROUGE set of measures assesses the difference between the generated summary and a collection of prior summaries. The ROUGE algorithm’s versions, ROUGE-1 (unigram overlap), ROUGE-2 (bigram overlap), and ROUGE-L (largest common subsequence), were employed to test accuracy, recall, and content overlap specifically.

1) ROUGE-1 (Unigram Overlap): This metric calculates the amount of overlap between the reference summaries and the generated summary for unigrams, or single words. It evaluates the recall as well as accuracy of unigrams, giving information on the level of content overlap among the references and the generated summary. Greater agreement between the generated and reference summaries in terms of individual words is indicated by higher ROUGE-1 scores.
2) ROUGE-2 (Bigram Overlap): This version of the evaluation expands to take into account bigram overlap—sequences of two consecutive words—between the reference summaries and the produced summary. ROUGE-2 offers a more comprehensive assessment of subject matter overlap than ROUGE-1 since it records the occurrence of word sequences. Better agreement between the produced and reference summaries’ word sequences is indicated by higher ROUGE-2 scores.
3) ROUGE-L (Longest Common Subsequence): ROUGE-L calculates the longest common subsequence. In contrast to ROUGE-1 and ROUGE-2, which emphasize on word matches exactly, ROUGE-L takes into consideration the longest continuous word sequence that shows up in both the reference summaries and the produced summary. This standard is particularly useful in determining the overall overlap of information and the degree of conceptual similarity between summaries, irrespective of word order.

Sample Instance: ROUGE scores calculated for a sample generated summary using the following metrics:

- ROUGE-1 (Unigram Overlap):
  1) Recall (R): 0.9
  2) Precision (P): 0.5625
  3) F1-score (F): 0.6923076875739645
- ROUGE-2 (Bigram Overlap):
  1) Recall (R): 0.7777777777777778
  2) Precision (P): 0.30434782608695654
  3) F1-score (F): 0.4374999959570314
- ROUGE-L (Longest Common Subsequence):
  1) Recall (R): 0.9
  2) Precision (P): 0.5625
  3) F1-score (F): 0.6923076875739645

Although ROUGE scores offer important insights into how effectively summarization models work, it’s essential to understand their limitations. ROUGE scores can fail to accurately represent the semantic equality between the generated summary and the reference summaries, which is one of their major drawbacks. They are also sensitive to semantic similarity. Furthermore, factors like coherence, readability, and fluency—all important variables in evaluating the general caliber of summaries—are not taken into consideration by ROUGE scores. Thus, in order to get an in-depth understanding of summary quality, ROUGE scores—while giving useful quantitative measurements—should be combined with qualitative evaluation and human judgment.

D. Fine-Tuning

In natural language processing (NLP), fine-tuning [15] is the process of altering a language model that was previously trained to fit a specific assignment or dataset by training it with data particular to the task. With this method, the model can gain task-specific patterns and complications while utilizing its previous expertise, which improves performance on the target job.

Motivation for Fine-Tuning: The need for adapting pre-trained models to particular datasets or tasks pushes fine-tuning. It’s possible that generic pre-trained models fail to perform at the highest level across every domain. Customization is made feasible by fine-tuning, which also guarantees that the model learns to produce outputs for the desired task or dataset that are more accurate and contextually appropriate.

Use of Fine-Tuning: Text classification, recognition of named entities, and text summarization are only a few of the NLP tasks that often make use of fine-tuning. It works particularly well in situations when the knowledge of the pre-trained model may be used but has to be modified to certain domains or situations.

How it helps in improving a good or service: By giving the model the ability to learn from specific to the task data, fine-tuning enhances the model's output. Through this process, the model's parameters are modified in order to better represent the nuanced characteristics and details of the target dataset, producing outputs that are both more precise and contextually relevant. Therefore, fine-tuning aids in improving the model's efficiency on the intended job, which eventually results in better summaries and overall outcomes.

IV. Results

A. Rouge Results

The performance of the summarization models differs across criteria, which can be shown by analyzing the ROUGE scores shown below in the Table: II. The ROUGE scores [10] were calculated through a comparison of the model-generated summaries to the dataset's reference summaries, using the ROUGE-1, ROUGE-2, and ROUGE-L standards. To ensure an accurate evaluation of the models' summarization ability, the average ROUGE scores for each model were calculated across the first 10 rows of the unknown dataset (SAMSum) [20]. This method offers a thorough evaluation of the model's effectiveness on several summarizing quality factors. MBZUAI/LaMini-Flan-T5-248M [5] is a standout model among those analyzed, having the best ROUGE ratings in nearly every parameter. This suggests that it is better than the reference summary at capturing unigram, bigram, and longest common subsequence overlaps. LaMini performs very well because of its strong architecture and efficient training process, which allow it to provide summaries that closely match the content of the original texts.

The MBZUAI/LaMini-Flan-T5-248M model [5] performed better than the others because of its greater size, fine-tuning process, and robust design. Because of its T5-based design, that allows for efficient encoding and decoding of textual data, it is well-suited for summarization work. Furthermore, the model may be strongly adjusted to the complexities of the task by performing fine-tuning using the LaMini-instruction dataset. With 248 million parameters, the model is bigger and has a higher representational ability. On the other hand, smaller model sizes, insufficient fine-tuning, or architectural limitations could have contributed to the lower scores of other models. When everything is considered, the MBZUAI/LaMini-Flan-T5-248M model [5] performs better than other models due to its strong architecture, efficient fine-tuning, and large model size, which allows it to produce summaries of excellent quality.

B. Fine tuning Results

Deciding to fine-tune the 'MBZUAI/LaMini-Flan-T5-248M' model [5] was based on its impressive performance in text summarizing tasks. A pre-trained model is easily fine-tuned to perform better on a particular task or dataset. Google Colab [21], a cloud-based platform, provided open access to GPU resources for model training. The Samsum dataset [20], which includes messenger-like interactions with summaries, was chosen for improving the model. The SAMSum dataset [20] was selected because it closely matches the conversational style and material of the model's initial training data, enabling the model to perform tasks that are more similar to each other. The overall goal of optimizing the 'MBZUAI/LaMini-Flan-T5-248M' model's [5] performance to generate excellent summaries in a conversational setting was accomplished via fine-tuning it on the Samsum dataset.

After choosing and fine-tuning the MBZUAI/LaMini-Flan-T5-248M model on the Samsum dataset, significant gains in its ROUGE scores were observed, as shown in Table III below. The model obtained ROUGE-1, ROUGE-2, and ROUGE-L scores of 37.83, 13.96 and 34.14, respectively, before fine-tuning. These scores improved to 41.29, 13.25, and 38.11, respectively, following adjustments. Hence, gains in recall, precision, and F1 scores across all three ROUGE versions demonstrate enhanced overlaps between the produced summaries and reference summaries. Overall, the MBZUAI/LaMini-Flan-T5-248M model's summarization abilities have improved significantly as a result of the fine-tuning procedure, increasing its efficacy in extracting important information from conversational text data. Furthermore, fine-tuning the MBZUAI/LaMini-Flan-T5-248M model [5] on the SAMSum dataset [20] highlights the potential of leveraging domain-specific data to optimize summarization performance.

V. Conclusion

In conclusion, the research examined how well different text summarizing algorithms performed side by side, providing insight into how well they captured the main ideas of the original texts. It was pointed out that models such as MBZUAI/LaMini-Flan-T5-248M displayed better summarization performance, with higher ROUGE scores on several criteria. Moreover, the significant enhancements in ROUGE scores following to the model's adaptation to the Samsum dataset demonstrated the clear impact of fine-tuning on model performance. This highlights how crucial customized training strategies are for maximizing model performance for certain tasks.

While automatic assessment measures like as ROUGE provide useful insights about summarization quality, they must be accompanied with human evaluation to identify details such as consistency and readability. With consideration of these factors, this dual evaluation technique guarantees a thorough assessment of summarization models.

Findings demonstrate how important text summarizing algorithms are for compressing large volumes of information into understandable summaries. These models provide efficiency and simplicity in processing textual data, with various number of uses ranging from content development to information retrieval. In the long run, text summarization research has a great deal of promise to enhance natural language processing and make information sharing easier in the digital era.

References

[1] "Hugging Face – The AI community building the future." Available: https://huggingface.co/

TABLE II
RESULTS OF ROUGE TEST

LLM Model	ROUGE-1			ROUGE-2			ROUGE-L		
	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
facebook/bart-large-cnn	35.43	20.19	25.04	9.53	5.40	6.70	34.26	19.73	24.39
Falconsai/text _summarization	23.54	16.66	18.90	7.17	5.31	5.81	20.95	14.77	16.79
MBZUAI/LaMini- Flan-T5-248M	51.61	30.71	37.83	19.94	11.19	13.96	45.97	27.95	34.14
sshleifer/distilbart -cnn-12-6	29.73	15.84	18.04	6.42	3.55	4.34	26.59	14.01	18.04
MBZUAI/LaMini-T5-61M	43.50	32.62	35.21	9.58	8.57	8.37	37.41	28.85	30.80
tuner007/pegasus _summarizer	31.99	18.82	23.39	6.94	3.95	5.02	30.96	18.13	22.56

TABLE III
FINE TUNING RESULTS

LLM Model	ROUGE-1			ROUGE-2			ROUGE-L		
	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
MBZUAI/LaMini- Flan-T5-248M	51.61	30.71	37.83	19.94	11.19	13.96	45.97	27.95	34.14
MBZUAI/LaMini-Flan-T5-248M (**Fine tuned**)	49.18	38.02	41.29	17.19	12.14	13.25	45.91	34.72	38.11

[2] "facebook/bart-large-cnn · Hugging Face," Jan. 18, 2001. Available: https://huggingface.co/facebook/bart-large-cnn

[3] M. Lewis et al., 'BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension', arXiv [cs.CL]. 2019. Available: https://arxiv.org/pdf/1910.13461.pdf

[4] "Falconsai/text_summarization · Hugging Face." Available: https://huggingface.co/Falconsai/text_summarization

[5] "MBZUAI/LAMINI-FLAN-T5-248M · Hugging face." Available: https://huggingface.co/MBZUAI/LaMini-Flan-T5-248M

[6] M. Wu, A. Waheed, C. Zhang, M. Abdul-Mageed, and A. F. Aji, 'LaMini-LM: A Diverse Herd of Distilled Models from Large-Scale Instructions', arXiv [cs.CL]. 2024. Available: https://arxiv.org/abs/2304.14402

[7] "sshleifer/distilbart-cnn-12-6 · Hugging Face," Jan. 18, 2001. Available: https://huggingface.co/sshleifer/distilbart-cnn-12-6

[8] "MBZUAI/LAMINI-T5-61M · Hugging face." Available: https://huggingface.co/MBZUAI/LaMini-T5-61M

[9] "tuner007/pegasus_summarizer · Hugging Face." Available: https://huggingface.co/tuner007/pegasus_summarizer

[10] "Metric Card for ROUGE, Hugging Face Metrics." [Online]. Available: https://huggingface.co/spaces/evaluate-metric/rouge

[11] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In Text Summarization Branches Out, pages 74–81, Barcelona, Spain. Association for Computational Linguistics. Available: https://aclanthology.org/W04-1013.pdf

[12] F. Chiusano, "Two minutes NLP — Learn the ROUGE metric by examples," Medium, Aug. 04, 2023. [Online]. Available: https://medium.com/nlplanet/two-minutes-nlp-learn-the-rouge-metric-by-examples-f179cc285499

[13] M. Mayo, "Approaches to text Summarization: An Overview - KDNuggets," KDnuggets. Avaiable: https://www.kdnuggets.com/2019/01/approaches-text-summarization-overview.html

[14] Cloudflare. "What Is a Large Language Model?" Cloudflare, Available: https://www.cloudflare.com/learning/ai/what-is-large-language-model/.

[15] "The Ultimate Guide to LLM Fine Tuning: Best Practices & Tools — Lakera – Protecting AI teams that disrupt the world." Available: https://www.lakera.ai/blog/llm-fine-tuning-guide

[16] R. Rastogi, "Papers explained 09: BART - DAIR.AI - medium," Medium, Sep. 20, 2023. [Online]. Available: https://medium.com/dair-ai/papers-explained-09-bart-7f56138175bd

[17] Q. Chen, "T5: a detailed explanation - Analytics Vidhya - Medium," Medium, Dec. 14, 2021. [Online]. Available: https://medium.com/analytics-vidhya/t5-a-detailed-explanation-a0ac9bc53e51

[18] Exemplary AI, "Accurate transcription and AI-Assisted editing & analysis," Exemplary AI, Feb. 02, 2023. Available: https://exemplary.ai/blog/flan-t5

[19] Shan, Ming-Chien. "Pegasus architecture and design principles." Proceedings of the 1993 ACM SIGMOD international conference on Management of data (1993): n. pag. Available: https://dl.acm.org/doi/pdf/10.1145/170036.170111

[20] B. Gliwa, I. Mochol, M. Biesek, and A. Wawer, 'SAMSum Corpus: A Human-annotated Dialogue Dataset for Abstractive Summarization', in Proceedings of the 2nd Workshop on New Frontiers in Summarization, 2019. Available: https://arxiv.org/pdf/1911.12237.pdf

[21] "Google Colaboratory." Available: https://colab.research.google.com/

Analysis of Transformer

by Manraj Singh

Submission date: 30-Mar-2024 08:42AM (UTC+0530)
Submission ID: 2335063061
File name: Analysis_of_Transformer.pdf (106.13K)
Word count: 4478
Character count: 27114

Analysis of Transformer based Pre-trained Summarization Models

Prof. Manisha Mathur
Assistant Professor, Computer Engineering
Vivekanand Education Society's Institute of Technology
Mumbai, India
manisha.mathur@ves.ac.in

Manraj Singh Virdi
Student, Computer Engineering
Vivekanand Education Society's Institute of Technology
Mumbai, India
d2021.manrajsingh.virdi@ves.ac.in

Nikhil Dhanwani
Student, Computer Engineering
Vivekanand Education Society's Institute of Technology
Mumbai, India
d2021.nikhil.dhanwani@ves.ac.in

Chirag Santwani
Student, Computer Engineering
Vivekanand Education Society's Institute of Technology
Mumbai, India
d2021.chirag.santwani@ves.ac.in

***Abstract*—In order to efficiently retrieve and comprehend large volumes of information, automated text summarization plays a crucial role in compressing the material into brief and understandable summaries. In this study, a thorough comparison of six innovative transformer-based summarization models sourced from the Hugging Face model hub is presented. Using ROUGE scores to evaluate sample outputs in both quantitative and qualitative terms, each model's performance and capabilities are carefully assessed, clarifying its unique advantages, disadvantages, and complexities. The results reveal significant variations in the models' summarization quality and effectiveness, likely stemming from differences in their architectures, pre-training sets, and fine-tuning techniques. The objective is to equip both students and professionals with the necessary knowledge to make informed decisions when selecting models for various summarization tasks, thereby laying the groundwork for future developments in transformer-based text summarization techniques.**

***Index Terms*—Automated text summarization, Transformer-based, ROUGE scores, Model performance, Pre-training sets, Fine-tuning techniques, Text summarization techniques.**

I. Introduction

In today's digital world, we can access a huge quantity of information with a few touches on our devices. This need more effective methods of understanding and managing all of this data. Text summary helps by carefully arranging and dividing enormous amounts of text into brief summaries while preserving crucial facts. Summarized text is even more crucial in addressing the information overload due to the necessity for rapid and efficient methods of finding information. Additionally, by reducing time and effort on information processing duties, text summarizing contributes to increased productivity. By giving brief summaries, it enables users to concentrate on key elements without getting distracted by information.

The field of deep learning and natural language processing has developed so rapidly that models that have been trained are now freely accessible and helpful for many NLP applications, especially text summarization. However, with so many models available—from BERT to T5—selecting the optimal pre-trained model for a specific job might be challenging. In addition, while pre-trained models offer an excellent place to begin, they can be even more effective and adaptable by modifying them with data specific to a certain area.

The motive of this exploration is to perform a thorough evaluation of six innovative text summarization models sourced from the Hugging Face model hub [1], considering the challenges posed by information overload. Specifically, the models examined include Facebook's BART-Large-CNN [2], Falconsai's Text-Summarization [4], MBZUAI's LaMini-Flan-T5-248M [5], sshleifer's DistilBART-CNN-12-6 [7], MBZUAI's LaMini-T5-61M [8], and Tuner007's Pegasus_summarizer [9]. This evaluation aims to highlight the advantages, disadvantages, and potential of each model, utilizing recognized metrics like ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores [10] and detailed examination of sample outcomes. Additionally, the investigation will explore the effectiveness of fine-tuning techniques in enhancing these models' performance on text summarization tasks. Through this comparative analysis, valuable insights can be provided to professionals and scholars, aiding them in navigating the landscape of summarized text models and optimizing their utility for managing information overload.

II. Background

Text summarization stands as a fundamental task within the domain of Natural Language Processing (NLP), addressing the challenge of condensing extensive volumes of textual data into succinct and coherent summaries. With an exponential surge in the creation of research papers, news articles, social media posts, and various digital materials, the need for effective summarization methodologies becomes increasingly paramount. Beyond mere condensation, text summarization serves as a facilitator for diverse tasks such as document classification, outcome ranking, and tailored content suggestion.

A. Approaches to Text Summarization

There are two primary approaches for text summary in natural language processing: extractive and abstractive summarization. In extractive summarizing [13], the precise wording and organization of the original information is maintained by picking out significant lines or paragraphs from the text as a whole and then presenting them in the form of a summary. In order to pick the most useful words based on factors including relevance, significance, and repetition, this method uses algorithms. Abstractive summary [13], on the other hand, involves developing new content that clearly and logically communicates the key points of the original text. Abstractive approaches clarify and paraphrase text using natural language comprehension and production processes, which frequently provide summaries that are more concise and logical.

B. Large Language Models (LLMs)

Among Large Language Models (LLMs) [14], Transformer-based models such as BART, GPT and T5 have transformed text summarization by their capacity to recognize complex language patterns and meaningful connections. This is because they have been pre-trained on huge amounts of text data. These models show exceptional understanding and generation skills for human-like writing, which makes them useful for a broad series of jobs with respect to natural language processing, including summarization. By using the encoded knowledge stored inside its boundaries, LLMs are able to efficiently compress complicated documents while maintaining crucial information and context. Analyzing automatically generated summaries is still difficult, though. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores [12] are frequently utilized metrics; however, they have several drawbacks, like their insensitivity to semantic similarity and their incapacity to capture readability and overall consistency. Thus, human assessment continues to be necessary for gaining an in-depth knowledge of summary quality, completing automated measures such as ROUGE and opens the door for future advances in the field.

C. Fine Tuning Models

In addition, optimizing Large Language Models (LLMs) on datasets customized to certain tasks offers an appealing method to improve their summarization performance. By significantly altering the pre-trained model's parameters, fine-tuning allows it to adapt its enormous prior knowledge to the specifics of a given summarization job or topic. It has been shown that this procedure significantly improves performance, helping the model to provide summaries that are more precise, relevant, and consistent. A crucial topic of investigation in this work is the possibility of fine-tuning LLMs for summarization tasks, especially since the goal is to expand the boundaries of the field by optimizing model performance on particular datasets.

III. Methodology

In this section, the thorough methods used to assess and compare the efficacy of six cutting-edge text summarization algorithms are described. The approach aims to evaluate how well each model generates concise and cohesive summaries in various domains. The criteria for model selection are outlined, along with the datasets utilized for evaluation and any fine-tuning processes. The evaluation metrics adopted to measure summarization quality are specified. To enhance flexibility and consistency of the results, the experimental environment, including system data and computational capabilities, is also provided. This technique attempts to investigate the possible advantages of fine-tuning to enhance model performance along with highlighting the relative benefits and drawbacks of each model. A significant contribution to the field of automated text summarizing is sought through the implementation of a systematic and rigorous evaluation approach that will direct further studies and their application innovation.

A. Model Selection

Text summarization tools capabilities and performance are significantly affected by their model architecture. Among the many different designs that are accessible, BART (Bidirectional and Auto-Regressive Transformers) [16] a transformer-based model noted for being flexible to natural language processing applications. It can efficiently collect context and produce coherent text because to its bidirectional architecture and autoregressive decoding. T5 (Text-To-Text Transfer Transformer) [17], designed by Google AI, uses a unified architecture in which the inputs and outputs are expressed as text strings, making it suitable for a variety of NLP applications. FLAN T5 [18] a version of T5 that has been specially optimized for text summary, providing greater efficiency in producing concise and insightful summaries. Another Google AI model, Pegasus [19], is well known for its abstractive summarization skills. It can efficiently extract important information from source texts and reword or rephrase it. Several models utilize these architectures, including 'facebook/bart-large-cnn' based on BART, 'MBZUAI/LaMini-T5-61M' based on T5, 'MBZUAI/LaMini-Flan-T5-248M' based on FLAN T5, and 'tuner007/pegasus_summarizer' based on Pegasus.

Selected Models

- **Model 1: facebook/bart-large-cnn**
  This model, as cited in [3], created by the research team at Facebook AI, is an effective tool for text summarization. It is based on the robust design of BART (Bidirectional and Auto-Regressive Transformers). The architecture allows the model to understand complicated language connections and structures observed in textual data. Designed specifically for news item overviews, 'facebook/BART-Large-CNN' is excellent at collecting important texts, which reduces the amount of information that is consumed. By carefully tuning with the CNN Daily Mail dataset—a comprehensive collection of text-summary pairs—the algorithm achieves outstanding results on summarizing tasks.
- **Model 2: Falconsai/text_summarization**

Designed by Falcon AI [4], it is aimed at the medical text summarizing industry. Based on a foundation pre-trained on a wide range of medical literature, this innovative approach makes use of an optimized version of the T5 Large transformer model. The model is excellent at producing concise and coherent summaries that are appropriate to the complexities of the medical area because it focuses on medical texts, such as research papers, clinical notes, and documents. By applying a thorough method of fine-tuning on a composite dataset gathered from various fields, the 'falconsai/text_summarization' model ensures its versatility and efficacy in a wide range of tasks.

- **Model 3: MBZUAI/LaMini-Flan-T5-248M**
  A refined variant of Google's Flan-T5-base model, called 'MBZUAI/LaMini-Flan-T5-248M' [6], is optimized to work with the LaMini-instruction dataset, which contains about 2.58 million instruction samples. It provides compact, effective language models with a range of sizes, checkpoints, and structures as part of the LaMini-LM series. In contrast to other models in the series, the 'MBZUAI/LaMini-Flan-T5-248M' utilizes an encoder-decoder architecture with 248 million parameters, that is based on the T5 model. Its adaptability and usability in a variety of environments are guaranteed by its rigorous evaluations across several NLP tasks and human assessments, demonstrating its flexibility to a wide range of natural language processing activities, including text generation.
- **Model 4: sshleifer/distilbart-cnn-12-6**
  The 'sshleifer DistilBART CNN 12-6' [7] model is a pre-trained language model designed for text summarization tasks based on the CNN/DailyMail dataset. It is a compressed and simplified variation of the BART model. Using a distillation process, DistilBART transfers the knowledge and performance of the bigger BART model by teaching a smaller student version to behave like it. Specifically designed for text summarization, the 'distilbart-cnn-12-6' variation uses a Convolutional Neural Network (CNN) encoder architecture with 12 layers of encoder and 6 decoder levels. This architecture improves the model's summarizing abilities by assisting in the acquisition of local contextual information in the input text.
- **Model 5: MBZUAI/LaMini-T5-61M**
  A key part of the LaMini-LM series, the 'MBZUAI/LaMini-T5-61M' [8] model was carefully developed by the Mohamed bin Zayed University of Artificial Intelligence (MBZUAI). Optimized for instruction fine-tuning, this version of the T5-small model has been developed on the LaMini-instruction dataset, an extensive collection of 2.58 million samples. The model has 61 million parameters, balances both efficiency and performance, which makes it a great choice for a range of natural language processing applications. For this study, the model chosen is 'MBZUAI/LaMini-T5-61M' because of its stable structure, tailored instruction using domain-specific data, and alignment with the goals stated in the research report "LaMini-LM: A Diverse Herd of Distilled Models from Large-Scale Instructions." [6]. This model is a good fit for summarization experiments because it provides a strong combination of computational effectiveness and task-specific expertise.
- **Model 6: tuner007/pegasus_summarizer**
  The pegasus_summarizer [9] is a text summarizing model that is available to the public and has been developed to maximize the potential of Google AI's Pegasus architecture. Since Pegasus serves as its foundation, it undoubtedly specializes at abstractive summarization and is skilled at utilizing various words and phrasing to express the main idea of the source material. A refined version of Pegasus, the 'tuner007/pegasus_summarizer' model is designed to improve performance on text summarizing work. It seeks to further improve summarization abilities by optimizing the original Pegasus model, giving consistency and informativeness in produced summaries.

TABLE I
OVERVIEW OF TEXT SUMMARIZATION MODELS

Model Name	**Description**	**Architecture**	**Size**
facebook/bart-large-cnn	BART with CNN encoder, fine-tuned on CNN Daily Mail dataset	Encoder-Decoder (BART) with CNN encoder	406M params
Falconsai/ text_summarization	Fine-tuned T5 Large for medical text summarization	Transformer (T5)	60.5M params
MBZUAI/LaMini-Flan-T5-248M	Fine-tuned Flan-T5-based on Lamini-instruction dataset	Transformer (Flan-T5)	248M params
sshleifer/distilbart-cnn-12-6	Distilled BART with CNN encoder for efficient summarization tasks	Encoder-Decoder (Distilled BART) with CNN encoder	306M params
MBZUAI/LaMini-T5-61M	Compact T5 variant for summarization	Transformer (T5)	61M params
tuner007/ pegasus_summarizer	An optimized model based on the powerful Pegasus architecture	Pegasus-based model	-

Table I provides a summary of the models examined in this study. After providing an overview of each model, the process of preparing the data is delved into, describing the dataset used for assessment and any processing measures taken to ensure consistency and precision in the evaluation of the model's performance.

B. Dataset

SAMSum Dataset: The SAMSum dataset [20] includes around 16,000 messenger-like conversations with summaries for each. These dialogues were created by English-speaking linguists who were given the task to create interactions that matched their regular talks. The goal of the dataset is to replicate the subject matter distribution seen in actual messenger chats. Discussions display a broad range of records and styles, including semi-formal, formal, and casual conversations. The conversations were then annotated with brief summaries that were meant to capture the key points addressed in the dialogues. The Samsung R&D Institute Poland created the SAMSum dataset, which is licensed under a non-commercial agreement (CC BY-NC-ND 4.0) for use in research.

Data Splits:

- Train: 14,732 conversations
- Validation: 818 conversations
- Test: 819 conversations

Sample Instance:

- "id": "13818513",
  "dialogue":
  - "Amanda: I baked cookies. Do you want some? Jerry: Sure! Amanda: I'll bring you some tomorrow."

  "summary":
  - "Amanda baked biscuits and will bring Jerry some hereafter",

C. Evaluation Strategy

In the research conducted, the main criteria utilized to evaluate the level of quality of the produced summaries was ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores [11]. The ROUGE set of measures assesses the difference between the generated summary and a collection of prior summaries. The ROUGE algorithm's versions, ROUGE-1 (unigram overlap), ROUGE-2 (bigram overlap), and ROUGE-L (largest common subsequence), were employed to test accuracy, recall, and content overlap specifically.

1) ROUGE-1 (Unigram Overlap): This metric calculates the amount of overlap between the reference summaries and the generated summary for unigrams, or single words. It evaluates the recall as well as accuracy of unigrams, giving information on the level of content overlap among the references and the generated summary. Greater agreement between the generated and reference summaries in terms of individual words is indicated by higher ROUGE-1 scores.
2) ROUGE-2 (Bigram Overlap): This version of the evaluation expands to take into account bigram overlap—sequences of two consecutive words—between the reference summaries and the produced summary. ROUGE-2 offers a more comprehensive assessment of subject matter overlap than ROUGE-1 since it records the occurrence of word sequences. Better agreement between the produced and reference summaries' word sequences is indicated by higher ROUGE-2 scores.
3) ROUGE-L (Longest Common Subsequence): ROUGE-L calculates the longest common subsequence. In contrast to ROUGE-1 and ROUGE-2, which emphasize on word matches exactly, ROUGE-L takes into consideration the longest continuous word sequence that shows up in both the reference summaries and the produced summary. This standard is particularly useful in determining the overall overlap of information and the degree of conceptual similarity between summaries, irrespective of word order.

Sample Instance: ROUGE scores calculated for a sample generated summary using the following metrics:

- ROUGE-1 (Unigram Overlap):
  1) Recall (R): 0.9
  2) Precision (P): 0.5625
  3) F1-score (F): 0.6923076875739645
- ROUGE-2 (Bigram Overlap):
  1) Recall (R): 0.7777777777777778
  2) Precision (P): 0.30434782608695654
  3) F1-score (F): 0.4374999959570314
- ROUGE-L (Longest Common Subsequence):
  1) Recall (R): 0.9
  2) Precision (P): 0.5625
  3) F1-score (F): 0.6923076875739645

Although ROUGE scores offer important insights into how effectively summarization models work, it's essential to understand their limitations. ROUGE scores can fail to accurately represent the semantic equality between the generated summary and the reference summaries, which is one of their major drawbacks. They are also sensitive to semantic similarity. Furthermore, factors like coherence, readability, and fluency—all important variables in evaluating the general caliber of summaries—are not taken into consideration by ROUGE scores. Thus, in order to get an in-depth understanding of summary quality, ROUGE scores—while giving useful quantitative measurements—should be combined with qualitative evaluation and human judgment.

D. Fine-Tuning

In natural language processing (NLP), fine-tuning [15] is the process of altering a language model that was previously trained to fit a specific assignment or dataset by training it with data particular to the task. With this method, the model can gain task-specific patterns and complications while utilizing its previous expertise, which improves performance on the target job.

Motivation for Fine-Tuning: The need for adapting pre-trained models to particular datasets or tasks pushes fine-tuning. It's possible that generic pre-trained models fail to perform at the highest level across every domain. Customization is made feasible by fine-tuning, which also guarantees that the model learns to produce outputs for the desired task or dataset that are more accurate and contextually appropriate.

Use of Fine-Tuning: Text classification, recognition of named entities, and text summarization are only a few of the NLP tasks that often make use of fine-tuning. It works particularly well in situations when the knowledge of the pre-trained model may be used but has to be modified to certain domains or situations.

How it helps in improving a good or service: By giving the model the ability to learn from specific to the task data, fine-tuning enhances the model's output. Through this process, the model's parameters are modified in order to better represent the nuanced characteristics and details of the target dataset, producing outputs that are both more precise and contextually relevant. Therefore, fine-tuning aids in improving the model's efficiency on the intended job, which eventually results in better summaries and overall outcomes.

IV. Results

A. Rouge Results

The performance of the summarization models differs across criteria, which can be shown by analyzing the ROUGE scores shown below in the Table: II. The ROUGE scores [10] were calculated through a comparison of the model-generated summaries to the dataset's reference summaries, using the ROUGE-1, ROUGE-2, and ROUGE-L standards. To ensure an accurate evaluation of the models' summarization ability, the average ROUGE scores for each model were calculated across the first 10 rows of the unknown dataset (SAMSum) [20]. This method offers a thorough evaluation of the model's effectiveness on several summarizing quality factors. MBZUAI/LaMini-Flan-T5-248M [5] is a standout model among those analyzed, having the best ROUGE ratings in nearly every parameter. This suggests that it is better than the reference summary at capturing unigram, bigram, and longest common subsequence overlaps. LaMini performs very well because of its strong architecture and efficient training process, which allow it to provide summaries that closely match the content of the original texts.

The MBZUAI/LaMini-Flan-T5-248M model [5] performed better than the others because of its greater size, fine-tuning process, and robust design. Because of its T5-based design, that allows for efficient encoding and decoding of textual data, it is well-suited for summarization work. Furthermore, the model may be strongly adjusted to the complexities of the task by performing fine-tuning using the LaMini-instruction dataset. With 248 million parameters, the model is bigger and has a higher representational ability. On the other hand, smaller model sizes, insufficient fine-tuning, or architectural limitations could have contributed to the lower scores of other models. When everything is considered, the MBZUAI/LaMini-Flan-T5-248M model [5] performs better than other models due to its strong architecture, efficient fine-tuning, and large model size, which allows it to produce summaries of excellent quality.

B. Fine tuning Results

Deciding to fine-tune the 'MBZUAI/LaMini-Flan-T5-248M' model [5] was based on its impressive performance in text summarizing tasks. A pre-trained model is easily fine-tuned to perform better on a particular task or dataset. Google Colab [21], a cloud-based platform, provided open access to GPU resources for model training. The Samsum dataset [20], which includes messenger-like interactions with summaries, was chosen for improving the model. The SAMSum dataset [20] was selected because it closely matches the conversational style and material of the model's initial training data, enabling the model to perform tasks that are more similar to each other. The overall goal of optimizing the 'MBZUAI/LaMini-Flan-T5-248M' model's [5] performance to generate excellent summaries in a conversational setting was accomplished via fine-tuning it on the Samsum dataset.

After choosing and fine-tuning the MBZUAI/LaMini-Flan-T5-248M model on the Samsum dataset, significant gains in its ROUGE scores were observed, as shown in Table III below. The model obtained ROUGE-1, ROUGE-2, and ROUGE-L scores of 37.83, 13.96 and 34.14, respectively, before fine-tuning. These scores improved to 41.29, 13.25, and 38.11, respectively, following adjustments. Hence, gains in recall, precision, and F1 scores across all three ROUGE versions demonstrate enhanced overlaps between the produced summaries and reference summaries. Overall, the MBZUAI/LaMini-Flan-T5-248M model's summarization abilities have improved significantly as a result of the fine-tuning procedure, increasing its efficacy in extracting important information from conversational text data. Furthermore, fine-tuning the MBZUAI/LaMini-Flan-T5-248M model [5] on the SAMSum dataset [20] highlights the potential of leveraging domain-specific data to optimize summarization performance.

V. Conclusion

In conclusion, the research examined how well different text summarizing algorithms performed side by side, providing insight into how well they captured the main ideas of the original texts. It was pointed out that models such as MBZUAI/LaMini-Flan-T5-248M displayed better summarization performance, with higher ROUGE scores on several criteria. Moreover, the significant enhancements in ROUGE scores following to the model's adaptation to the Samsum dataset demonstrated the clear impact of fine-tuning on model performance. This highlights how crucial customized training strategies are for maximizing model performance for certain tasks.

While automatic assessment measures like as ROUGE provide useful insights about summarization quality, they must be accompanied with human evaluation to identify details such as consistency and readability. With consideration of these factors, this dual evaluation technique guarantees a thorough assessment of summarization models.

Findings demonstrate how important text summarizing algorithms are for compressing large volumes of information into understandable summaries. These models provide efficiency and simplicity in processing textual data, with various number of uses ranging from content development to information retrieval. In the long run, text summarization research has a great deal of promise to enhance natural language processing and make information sharing easier in the digital era.

References

[1] "Hugging Face – The AI community building the future." Available: https://huggingface.co/

TABLE II
RESULTS OF ROUGE TEST

LLM Model	ROUGE-1			ROUGE-2			ROUGE-L		
	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
facebook/bart-large-cnn	35.43	20.19	25.04	9.53	5.40	6.70	34.26	19.73	24.39
Falconsai/text _summarization	23.54	16.66	18.90	7.17	5.31	5.81	20.95	14.77	16.79
MBZUAI/LaMini- Flan-T5-248M	51.61	30.71	37.83	19.94	11.19	13.96	45.97	27.95	34.14
sshleifer/distilbart -cnn-12-6	29.73	15.84	18.04	6.42	3.55	4.34	26.59	14.01	18.04
MBZUAI/LaMini-T5-61M	43.50	32.62	35.21	9.58	8.57	8.37	37.41	28.85	30.80
tuner007/pegasus _summarizer	31.99	18.82	23.39	6.94	3.95	5.02	30.96	18.13	22.56

TABLE III
FINE TUNING RESULTS

LLM Model	ROUGE-1			ROUGE-2			ROUGE-L		
	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
MBZUAI/LaMini- Flan-T5-248M	51.61	30.71	37.83	19.94	11.19	13.96	45.97	27.95	34.14
MBZUAI/LaMini-Flan-T5-248M **(Fine tuned)**	49.18	38.02	41.29	17.19	12.14	13.25	45.91	34.72	38.11

[2] "facebook/bart-large-cnn · Hugging Face," Jan. 18, 2001. Available: https://huggingface.co/facebook/bart-large-cnn

[3] M. Lewis et al., 'BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension', arXiv [cs.CL]. 2019. Available: https://arxiv.org/pdf/1910.13461.pdf

[4] "Falconsai/text_summarization · Hugging Face." Available: https://huggingface.co/Falconsai/text_summarization

[5] "MBZUAI/LAMINI-FLAN-T5-248M · Hugging face." Available: https://huggingface.co/MBZUAI/LaMini-Flan-T5-248M

[6] M. Wu, A. Waheed, C. Zhang, M. Abdul-Mageed, and A. F. Aji, 'LaMini-LM: A Diverse Herd of Distilled Models from Large-Scale Instructions', arXiv [cs.CL]. 2024. Available: https://arxiv.org/abs/2304.14402

[7] "sshleifer/distilbart-cnn-12-6 · Hugging Face," Jan. 18, 2001. Available: https://huggingface.co/sshleifer/distilbart-cnn-12-6

[8] "MBZUAI/LAMINI-T5-61M · Hugging face." Available: https://huggingface.co/MBZUAI/LaMini-T5-61M

[9] "tuner007/pegasus_summarizer · Hugging Face." Available: https://huggingface.co/tuner007/pegasus_summarizer

[10] "Metric Card for ROUGE, Hugging Face Metrics." [Online]. Available: https://huggingface.co/spaces/evaluate-metric/rouge

[11] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In Text Summarization Branches Out, pages 74–81, Barcelona, Spain. Association for Computational Linguistics. Available: https://aclanthology.org/W04-1013.pdf

[12] F. Chiusano, "Two minutes NLP — Learn the ROUGE metric by examples," Medium, Aug. 04, 2023. [Online]. Available: https://medium.com/nlplanet/two-minutes-nlp-learn-the-rouge-metric-by-examples-f179cc285499

[13] M. Mayo, "Approaches to text Summarization: An Overview - KDNuggets," KDnuggets. Avaiable: https://www.kdnuggets.com/2019/01/approaches-text-summarization-overview.html

[14] Cloudflare. "What Is a Large Language Model?" Cloudflare, Available: https://www.cloudflare.com/learning/ai/what-is-large-language-model/.

[15] "The Ultimate Guide to LLM Fine Tuning: Best Practices & Tools — Lakera – Protecting AI teams that disrupt the world." Available: https://www.lakera.ai/blog/llm-fine-tuning-guide

[16] R. Rastogi, "Papers explained 09: BART - DAIR.AI - medium," Medium, Sep. 20, 2023. [Online]. Available: https://medium.com/dair-ai/papers-explained-09-bart-7f56138175bd

[17] Q. Chen, "T5: a detailed explanation - Analytics Vidhya - Medium," Medium, Dec. 14, 2021. [Online]. Available: https://medium.com/analytics-vidhya/t5-a-detailed-explanation-a0ac9bc53e51

[18] Exemplary AI, "Accurate transcription and AI-Assisted editing & analysis," Exemplary AI, Feb. 02, 2023. Available: https://exemplary.ai/blog/flan-t5

[19] Shan, Ming-Chien. "Pegasus architecture and design principles." Proceedings of the 1993 ACM SIGMOD international conference on Management of data (1993): n. pag. Available: https://dl.acm.org/doi/pdf/10.1145/170036.170111

[20] B. Gliwa, I. Mochol, M. Biesek, and A. Wawer, 'SAMSum Corpus: A Human-annotated Dialogue Dataset for Abstractive Summarization', in Proceedings of the 2nd Workshop on New Frontiers in Summarization, 2019. Available: https://arxiv.org/pdf/1911.12237.pdf

[21] "Google Colaboratory." Available: https://colab.research.google.com/

Analysis of Transformer

ORIGINALITY REPORT

3%	2%	1%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	d197for5662m48.cloudfront.net Internet Source	1%
2	Submitted to Liverpool John Moores University Student Paper	1%
3	Submitted to Sardar Patel College of Engineering Student Paper	1%
4	deepai.org Internet Source	1%

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		